Journal Of Philosophy
Vol. 88 1991 No.1-9

G. K. V.
HARDWAR

110146

110146

RS-1081

THE JOURNAL OF PHILOSOPHY

110146

११३४९

पुस्तकालय
गुरुकुल कांगड़ी विश्वविद्यालय
हरिद्वार

VOLUME LXXXVIII, NUMBER 1
JANUARY 1991

Vol. 88
1991 No 1–9

page
5 *A Modal Argument for Narrow Content* Jerry A. Fodor
27 *Real Patterns* Daniel C. Dennett

BOOK REVIEWS
52 Arthur W. Collins: *The Nature of Mental Things* Anita Avramides

Published by The Journal of Philosophy, Inc.

THE JOURNAL OF PHILOSOPHY
FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Arthur C. Danto, Kent Greenawalt, Charles Larmore, Isaac Levi, Mary Mothersill. *Special Projects Editor:* Sidney Morgenbesser. *Consulting Editors:* Akeel Bilgrami, James T. Higginbotham, Shaughan Lavine, Charles D. Parsons. *Managing Editor:* Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by The Journal of Philosophy, Inc. *President,* Arthur C. Danto; *Vice President,* Corliss Lamont; *Secretary,* Michael J. Mooney; *Treasurer,* Barbara Gimbel; *Other Trustees,* Leigh S. Cauman, Kent Greenawalt, André Schiffrin.

All communications to the editors and trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, 709 Philosophy Hall, Columbia University, New York, NY 10027. FAX: (212) 932-3721

THE JOURNAL OF PHILOSOPHY

SUBSCRIPTIONS (12 issues)

Individuals	$30.00
Libraries and Institutions	$50.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$ 8.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; printed by Lancaster Press, Inc., Lancaster, PA. Typeset by TAPSCO, Inc., Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, 709 Philosophy Hall, Columbia University, New York, New York, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL'S terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1991 by the Journal of Philosophy, Inc.

ISSN 0022-362X
POSTMASTER: Second class postage paid at New York, NY, and additional mailing offices. Send address changes to *Journal of Philosophy,* 709 Philosophy, Columbia U, NYC 10027.

THE JOURNAL OF PHILOSOPHY

25591

पुस्तकालय
गुरुकुल कांगड़ी विश्वविद्यालय
हरिद्वार

VOLUME LXXXVIII, NUMBER 3
MARCH 1991

page
113 *Epistemic Paternalism: Communication Control in Law and Society* Alvin I. Goldman
132 *Epistemology on Holiday* Mark Kaplan

BOOK REVIEWS
155 Jeremy Waldron: *The Right to Private Property* Alan Ryan
159 Richard W. Miller: *Fact and Method* Peter Kosso
163 Ronald N. Giere: *Explaining Science: A Cognitive Approach* Philip Kitcher

167 NOTES AND NEWS

Published by The Journal of Philosophy, Inc.

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Arthur C. Danto, Kent Greenawalt, Charles Larmore, Isaac Levi, Mary Mothersill. *Special Projects Editor:* Sidney Morgenbesser. *Consulting Editors:* Akeel Bilgrami, James T. Higginbotham, Shaughan Lavine, Charles D. Parsons. *Managing Editor:* Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by The Journal of Philosophy, Inc. *President,* Arthur C. Danto; *Vice President,* Corliss Lamont; *Secretary,* Michael J. Mooney; *Treasurer,* Barbara Gimbel; *Other Trustees,* Leigh S. Cauman, Kent Greenawalt, André Schiffrin.

All communications to the editors and trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, 709 Philosophy Hall, Columbia University, New York, NY 10027. FAX: (212) 932-3721

THE JOURNAL OF PHILOSOPHY

SUBSCRIPTIONS (12 issues)

Individuals	$30.00
Libraries and Institutions	$50.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$ 8.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; printed by Lancaster Press, Inc., Lancaster, PA. Typeset by TAPSCO, Inc., Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, 709 Philosophy Hall, Columbia University, New York, New York, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL'S terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1991 by the Journal of Philosophy, Inc.

ISSN 0022-362X

POSTMASTER: Second class postage paid at New York, NY, and additional mailing offices. Send address changes to *Journal of Philosophy,* 709 Philosophy, Columbia U, NYC 10027.

THE JOURNAL OF PHILOSOPHY

25|5|91

VOLUME LXXXVIII, NUMBER 4
APRIL 1991

page
169 *How Euclidean Geometry Has Misled Metaphysics* Graham Nerlich
190 *Agency and Morality* Richard Brook

BOOK REVIEWS
213 Paul Grice: *Studies in the Way of Words* Robert J. Fogelin
219 Paul Feyerabend: *Against Method & Farewell to Reason* Ian Hacking

223 NEW BOOKS: ANTHOLOGIES

Published by The Journal of Philosophy, Inc.

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Arthur C. Danto, Kent Greenawalt, Charles Larmore, Isaac Levi, Mary Mothersill. *Special Projects Editor:* Sidney Morgenbesser. *Consulting Editors:* Akeel Bilgrami, James T. Higginbotham, Shaughan Lavine, Charles D. Parsons. *Managing Editor:* Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by The Journal of Philosophy, Inc. *President,* Arthur C. Danto; *Vice President,* Corliss Lamont; *Secretary,* Michael J. Mooney; *Treasurer,* Barbara Gimbel; *Other Trustees,* Leigh S. Cauman, Kent Greenawalt, André Schiffrin.

All communications to the editors and trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, 709 Philosophy Hall, Columbia University, New York, NY 10027. FAX: (212) 932-3721

THE JOURNAL OF PHILOSOPHY

SUBSCRIPTIONS (12 issues)

Individuals	$30.00
Libraries and Institutions	$50.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$ 8.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; printed by Lancaster Press, Inc., Lancaster, PA. Typeset by TAPSCO, Inc., Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, 709 Philosophy Hall, Columbia University, New York, New York, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL'S terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1991 by the Journal of Philosophy, Inc.

ISSN 0022-362X

POSTMASTER: Second class postage paid at New York, NY, and additional mailing offices. Send address changes to *Journal of Philosophy,* 709 Philosophy, Columbia U, NYC 10027.

THE JOURNAL OF PHILOSOPHY

121791

VOLUME LXXXVIII, NUMBER 5
MAY 1991

page		
225	*The What and the How*	Joseph Almog
245	*'P, therefore, P' without Circularity*	Roy A. Sorensen
	BOOK REVIEWS	
267	G.A. Cohen: *History, Labour and Freedom: Themes from Marx*	Andrew Levine
275	Arthur Fine: *The Shaky Game: Einstein, Realism, and the Quantum Theory*	R. I. G. Hughes
280	NEW BOOKS: TRANSLATIONS	

Published by The Journal of Philosophy, Inc.

THE JOURNAL OF PHILOSOPHY
FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Arthur C. Danto, Kent Greenawalt, Charles Larmore, Isaac Levi, Mary Mothersill. *Special Projects Editor:* Sidney Morgenbesser. *Consulting Editors:* Akeel Bilgrami, James T. Higginbotham, Shaughan Lavine, Charles D. Parsons. *Managing Editor:* Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by The Journal of Philosophy, Inc. *President,* Arthur C. Danto: *Vice President,* Corliss Lamont; *Secretary,* Michael J. Mooney; *Treasurer,* Barbara Gimbel; *Other Trustees,* Leigh S. Cauman, Kent Greenawalt, André Schiffrin.

All communications to the editors and trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, 709 Philosophy Hall, Columbia University, New York, NY 10027. FAX: (212) 932-3721.

THE JOURNAL OF PHILOSOPHY

SUBSCRIPTIONS (12 issues)

Individuals	$30.00
Libraries and Institutions	$50.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$ 8.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; printed by Lancaster Press, Inc., Lancaster, PA. Typeset by TAPSCO, Inc., Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, 709 Philosophy Hall, Columbia University, New York, New York, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1991 by the Journal of Philosophy, Inc. ISSN 0022-362X

POSTMASTER: Second class postage paid at New York, NY, and additional mailing offices. Send address changes to *Journal of Philosophy,* 709 Philosophy, Columbia U, NYC 10027.

THE JOURNAL OF PHILOSOPHY

17/7/91

VOLUME LXXXVIII, NUMBER 6
JUNE 1991

page
281 *Contractualism, Moral Motivation, and Practical Reason* Samuel Freeman

COMMENTS AND CRITICISM

304 *The Causal History of Computational Activity: Maudlin and Olympia* Eric Barnes

BOOK REVIEWS

317 Stuart Hampshire: *Innocence and Experience* David Luban
324 Eva Kittay: *Metaphor: Its Cognitive Force and Linguistic Structure* Peter Ludlow
331 Michael Detlefsen: *Hilbert's Program: An Essay on Mathematical Instrumentalism* Mark Steiner

Published by The Journal of Philosophy, Inc.

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Arthur C. Danto, Kent Greenawalt, Charles Larmore, Isaac Levi, Mary Mothersill. *Special Projects Editor:* Sidney Morgenbesser. *Consulting Editors:* Akeel Bilgrami, James T. Higginbotham, Shaughan Lavine, Charles D. Parsons. *Managing Editor:* Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by The Journal of Philosophy, Inc. *President,* Arthur C. Danto: *Vice President,* Corliss Lamont; *Secretary,* Michael J. Mooney; *Treasurer,* Barbara Gimbel; *Other Trustees,* Leigh S. Cauman, Kent Greenawalt, André Schiffrin.

All communications to the editors and trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, 709 Philosophy Hall, Columbia University, New York, NY 10027. FAX: (212) 932-3721.

THE JOURNAL OF PHILOSOPHY

SUBSCRIPTIONS (12 issues)

Individuals	$30.00
Libraries and Institutions	$50.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$ 8.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; printed by Lancaster Press, Inc., Lancaster, PA. Typeset by TAPSCO, Inc., Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, 709 Philosophy Hall, Columbia University, New York, New York, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1991 by the Journal of Philosophy, Inc.

ISSN 0022-362X

POSTMASTER: Second class postage paid at New York, NY, and additional mailing offices. Send address changes to *Journal of Philosophy,* 709 Philosophy, Columbia U, NYC 10027.

THE JOURNAL OF PHILOSOPHY

11819

VOLUME LXXXVIII, NUMBER 7
JULY 1991

page
337 *Reductive and Nonreductive Simile Theories of Metaphor* Lynne Tirrell
359 *Reference and Pronominal Descriptions* George M. Wilson
388 NEW BOOKS

Published by The Journal of Philosophy, Inc.

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Arthur C. Danto, Kent Greenawalt, Charles Larmore, Isaac Levi, Mary Mothersill. *Special Projects Editor:* Sidney Morgenbesser. *Consulting Editors:* Akeel Bilgrami, James T. Higginbotham, Shaughan Lavine, Charles D. Parsons. *Managing Editor:* Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by The Journal of Philosophy, Inc. *President,* Arthur C. Danto: *Vice President,* Corliss Lamont; *Secretary,* Michael J. Mooney; *Treasurer,* Barbara Gimbel; *Other Trustees,* Leigh S. Cauman, Kent Greenawalt, André Schiffrin.

All communications to the editors and trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, 709 Philosophy Hall, Columbia University, New York, NY 10027. FAX: (212) 932-3721.

THE JOURNAL OF PHILOSOPHY

SUBSCRIPTIONS (12 issues)

Individuals	$30.00
Libraries and Institutions	$50.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$ 8.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; printed by Lancaster Press, Inc., Lancaster, PA. Typeset by TAPSCO, Inc., Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, 709 Philosophy Hall, Columbia University, New York, New York, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1991 by the Journal of Philosophy, Inc.

ISSN 0022-362X

POSTMASTER: Second class postage paid at New York, NY, and additional mailing offices.

THE JOURNAL OF PHILOSOPHY

VOLUME LXXXVIII, NUMBER 8
AUGUST 1991

page
393 *Holism and Nonseparability* Richard A. Healey

BOOK REVIEWS
422 Charles Taylor: *Sources of the Self* J. B. Schneewind
427 William Hasker: *God, Time, and Knowledge* and Edward R. Wierenga: *The Nature of God: An Inquiry into Divine Attributes* John Martin Fischer
434 Judea Pearl: *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* Henry E. Kyburg, Jr.
437 Peter Gärdenfors: *Knowledge in Flux: Modeling the Dynamics of Epistemic States* Isaac Levi

445 New Books: Anthologies
446 New Books: Paperbacks
447 New Books: Translations

Published by The Journal of Philosophy, Inc.

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Arthur C. Danto, Kent Greenawalt, Charles Larmore, Isaac Levi, Mary Mothersill. *Special Projects Editor:* Sidney Morgenbesser. *Consulting Editors:* Akeel Bilgrami, James T. Higginbotham, Shaughan Lavine, Charles D. Parsons. *Managing Editor:* Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by The Journal of Philosophy, Inc. *President,* Arthur C. Danto: *Vice President,* Corliss Lamont; *Secretary,* Michael J. Mooney; *Treasurer,* Barbara Gimbel; *Other Trustees,* Leigh S. Cauman, Kent Greenawalt, André Schiffrin.

All communications to the editors and trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, 709 Philosophy Hall, Columbia University, New York, NY 10027. FAX: (212) 932-3721.

THE JOURNAL OF PHILOSOPHY

SUBSCRIPTIONS (12 issues)

Individuals	$30.00
Libraries and Institutions	$50.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$ 8.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; printed by Lancaster Press, Inc., Lancaster, PA. Typeset by TAPSCO, Inc., Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, 709 Philosophy Hall, Columbia University, New York, New York, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1991 by the Journal of Philosophy, Inc.

ISSN 0022-362X

POSTMASTER: Second class postage paid at New York, NY, and additional mailing offices. Send address changes to *Journal of Philosophy,* 709 Philosophy, Columbia U, NYC 10027.

THE JOURNAL OF PHILOSOPHY

20/10.91

पुस्तकालय
विश्वविद्यालय
हरिद्वार

VOLUME LXXXVIII, NUMBER 9
SEPTEMBER 1991

page
449 *Empirical Equivalence and Underdetermination* Larry Laudan and Jarrett Leplin
473 *Penny Pinching and Backward Induction* Martin Hollis
BOOK REVIEWS
489 T. H. Irwin: *Aristotle's First Principles* Gisela Striker
497 Rüdiger Bittner: *What Reason Demands* Thomas E. Hill, Jr.
501 NOTES AND NEWS
502 NEW BOOKS

Published by The Journal of Philosophy, Inc.

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Arthur C. Danto, Kent Greenawalt, Charles Larmore, Isaac Levi, Mary Mothersill. *Special Projects Editor:* Sidney Morgenbesser. *Consulting Editors:* Akeel Bilgrami, James T. Higginbotham, Shaughan Lavine, Charles D. Parsons. *Managing Editor:* Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by The Journal of Philosophy, Inc. *President,* Arthur C. Danto: *Vice President,* Corliss Lamont; *Secretary,* Michael J. Mooney; *Treasurer,* Barbara Gimbel; *Other Trustees,* Leigh S. Cauman, Kent Greenawalt, André Schiffrin.

All communications to the editors and trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, 709 Philosophy Hall, Columbia University, New York, NY 10027. FAX: (212) 932-3721.

THE JOURNAL OF PHILOSOPHY

SUBSCRIPTIONS (12 issues)

Individuals	$30.00
Libraries and Institutions	$50.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$ 8.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; printed by Lancaster Press, Inc., Lancaster, PA. Typeset by TAPSCO, Inc., Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, 709 Philosophy Hall, Columbia University, New York, New York, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1991 by the Journal of Philosophy, Inc.

ISSN 0022-362X

POSTMASTER: Second class postage paid at New York, NY, and additional mailing offices. Send address changes to *Journal of Philosophy,* 709 Philosophy, Columbia U, NYC 10027.

THE JOURNAL OF
PHILOSOPHY

VOLUME LXXXVIII
JANUARY–DECEMBER 1991

Vol. 88
1991
No1- 9*

Edited by
Bernard Berofsky, Arthur C. Danto
Kent Greenawalt, Charles Larmore
Isaac Levi, Sidney Morgenbesser
Mary Mothersill

Published by The Journal of Philosophy, Inc.

THE JOURNAL OF PHILOSOPHY was founded in 1904 as THE JOURNAL OF PHILOSOPHY, PSYCHOLOGY AND SCIENTIFIC METHODS by Frederick J. E. Woodbridge and J. McKeen Cattell. In 1906 Wendell T. Bush became associated with the JOURNAL as co-editor. In 1923, the JOURNAL was incorporated in the State of New York under its present name. Since its founding it has been published from Columbia University and printed by the Lancaster Press in Lancaster, Pennsylvania.

THE JOURNAL OF PHILOSOPHY

STATEMENT OF COPYRIGHT POLICY

We would like to draw your attention to the protection copyright affords against multiple photocopying by libraries, instructors, and copy centers. The 1976 copyright law provides that photocopies in excess of "fair use" copying cannot be made without authorization by the copyright owner—"fair use" meaning single copying for personal use. Accordingly, the *Journal* authorizes personal or educational multiple copying only if a request is made in writing, and the fee of $.05 per page/per copy is paid directly to us. *Each copy must include a notice of copyright.* We will turn over to the authors any substantial payments (over $20.00, an administrative fee) obtained from all photocopying. Inquiries about photocopying may be sent to the Managing Editor.

© Copyright 1991 by the Journal of Philosophy, Inc.

ISSN 0022-362X

POSTMASTER: Second class postage paid at New York, NY, and additional mailing offices.
Send address changes to *Journal of Philosophy*, 709 Philosophy, Columbia U, NYC 10027.

THE JOURNAL OF PHILOSOPHY

VOLUME LXXXVIII, NO. 1, JANUARY 1991

A MODAL ARGUMENT FOR NARROW CONTENT*

HERE is a modern antinomy. On the one hand, there is argument *A*:

Argument *A*:

1. My twin and I are molecular duplicates.
2. Therefore our (actual and counterfactual) behaviors are identical in relevant respects.
3. Therefore the causal powers of our mental states are identical in relevant respects.
4. Therefore my twin and I belong to the same natural kind for purposes of psychological explanation and "individualism" is true.

But, on the other hand, there is argument *B*:

Argument *B*:

1′ My twin and I are molecular duplicates.
2′ Nevertheless, our (actual and counterfactual) behaviors are different in relevant respects.
3′ Therefore the causal powers of our mental states are different in relevant respects.
4′ Therefore my twin and I belong to different natural kinds for purposes of psychological explanation and "individualism" is false.

At least one of these arguments must be unsound. Which one? And what is wrong with it?

In chapter 2 of *Psychosemantics*,[1] I offered some considerations intended to advance the cause of argument *A*. They were supposed

---

* This paper is deeply indebted to conversations with Georges Rey and Steve Stich. Help from Ned Block, Anne Jacobson, Tim Maudlin, Colin McGinn, Brian McLaughlin, and Stephen Schiffer is also gratefully acknowledged, as are edifying emails from Fred Adams, Joe Levine, and David Rosenthal.

[1] Cambridge: MIT, 1987.

0022-362X/91/8801/5-26

to show that mental states that differ only in "broad" intentional properties (the sorts of intentional properties that the mental states of molecular twins may fail to share) ipso facto do not differ in causal powers; hence that mere differences in broad intentional content do not determine differences in natural kinds for purposes of psychological explanation. The arguments I offered in *Psychosemantics* were not, however, greeted with unequivocal enthusiasm. Such was their subtlety, in fact, that I am not, myself, always quite sure how they were supposed to go. In the present paper, I propose to have another try at bolstering argument *A*. I shall use many of the same materials that I did in *Psychosemantics,* but I shall put the pieces together somewhat differently. In passing I shall have a word or two to say about some of the comments that the *Psychosemantics* arguments have provoked.

A preliminary remark, however, before we are under way. For most philosophical purposes, it may not really matter much how the individualism issue turns out. For example, I do not think that a resolution in favor of argument *A* (that is, in favor of individualism) would affect the status of "externalism" in semantics. Externalism is independent of individualism because, whatever the *explanatory* status of broad content, it is not in dispute (anyhow, it is not in *this* dispute) that the content of my twin's *water* thoughts differs from the content of mine; or that 'water' means something different in my mouth and in his; or that these semantical differences derive from differences in our respective head/world relations. (Presumably they derive from the fact that, whereas the causal history of my *water* thoughts connects them to samples of H_2O, the causal history of his connects them to samples of *XYZ*). If, in short, the intuitions about twins make a case for content externalism at all, then that case stands whether or not broad intentional states determine natural kinds for the purposes of causal explanation in psychology.

Similarly—and pace, for example, Steven Stich[2]—it is by no means obvious that the scientific vindication of intentional realism depends on how the issues about individualism are resolved. Suppose that there is some metaphysical argument to show that causal explanation in psychology requires the individuation of mental states to be individualistic (so that the mental states of molecular twins are ipso facto type-identical, as in argument *A*.) It would follow that psychological explanation is not, strictly speaking, a species of belief/desire explanation since it is common ground that beliefs and desires are individuated broadly. But it would still be open whether psychological explanations are species of *intentional* explanations. That would depend on whether an individualistic notion of "intentional state" suitable for the purposes of psychological explanation can be constructed. (For more on this, see *Psychosemantics*.)

---

[2] *From Folk Psychology to Cognitive Science* (Cambridge: MIT, 1983).

In short, you can do quite a lot of business in semantics and the philosophy of mind while benignly neglecting the issues about individualism. Still, I think it is worth trying to get these issues straight. We are about to see that they raise some interesting questions about the family of notions that include causal explanation, causal power, type identity, and natural kind. Getting some of this stuff sorted out may therefore be of use to metaphysics and the philosophy of science, even if it leaves things more or less unaltered in semantics and the philosophy of mind. So here we go.

Argument *A* says that, in virtue of our molecular identity, my twin's behavior and my behavior are identical *in all relevant respects*. Not, of course, that they are identical *tout court*. On the contrary, just as it is common ground in this discussion that the *mental states* of twins can differ in certain of their intentional properties, so it is common ground that twin *behaviors* can differ under some of their intentional descriptions. Indeed, it may be that the second concession is entailed by the first, since it is plausible that the intentional properties of behaviors are inherited from the intentional contents of their mental causes. Thus, it is plausibly because I can think about water and my twin cannot that I but not my twin can reach for water, drill for water, or recommend water to a thirsty friend. A committed individualist might wish to argue that none of these is really, strictu dictu, an example of *behavioral* description; but that is not a tack that I propose to take.

So then: twins can differ in the contents of their mental states; and, in consequence of these mental state differences, the behaviors of twins can differ in certain of their intentional properties. I shall take for granted (what is not, however, uncontested) that these differences in intentional properties are the *only* (relevant) differences between the twins' behaviors. That is, I am taking it for granted that the behaviors of twins are ipso facto identical in all nonintentional properties that are relevant to psychological taxonomy. That is, of course, stronger than assuming what, I think, literally everybody agrees about, namely, that our behaviors are identical under *physical* descriptions, i.e., identical qua motions.

Here is an apparent exception to the stronger claim: if I utter 'Gimme water', then, if all goes well, I get water; but if my twin utters 'Gimme water', then, if all goes well, he gets twater. So maybe there are nonintentional descriptions under which our behaviors are relevantly different after all. Second thoughts, however, make this difference go away; it is an artifact of the boring consideration that, whereas my utterance happened on earth, his happened on twin-earth. The salient consideration is that, if twin-me had uttered 'Gimme water' here, he would have gotten water; and if I had uttered 'Gimme water' on twin-earth, I would have gotten twater. To put it slightly differently, my 'Gimme water' does not get water *come what may;* at best it gets water *in certain circumstances*. But in *those* circumstances, his 'Gimme water' gets water, too.[3]

The moral is that you have to judge identity and difference of causal powers in a way that bears the counterfactuals in mind, namely, *across* contexts rather than *within* contexts.[4] That does not get us out of the woods, however. For notice that, in the case where I utter 'Gimme water' on Earth$_2$ and my twin utters it here, neither of us gets what he asks for. That is, whatever the context of utterance, my utterance is a water request and his utterance is a twater request. So our behaviors remain relevantly different under these intentional descriptions *even by the across-context test*. It is this residual difference between the behaviors—their cross-context difference under certain intentional descriptions—which is the challenge to individualism and local supervenience.

---

[3] Here is a related case—suggested by an example of Colin McGinn's—that I take to be susceptible to the same sort of treatment. Suppose on earth$_2$, there is not only twin-water, but also twin-salt (it is LCaN rather than NaCl). And suppose the fact is that thinking about salt makes you want water (whereas thinking about LCaN makes your twin want *XYZ*). So, would there not then be a difference between the causal powers of salt thoughts and twin-salt thoughts in virtue of these differences between the contents of the wants that they cause? No, because salt thoughts do not have the power to cause water wants *come what may;* what they have is the power to cause water wants *in somebody who has the concept water*. (That is, in somebody who has the appropriate causal/historical—or whatever—connections to H_2O.) But, of course, twin-salt thoughts have *that* power too (just as salt thoughts have the power to cause twater wants in somebody who has the appropriate causal/historical connections to *XYZ*.) Here as elsewhere, one applies the cross-context test by asking whether *A* would have the same effects as *B* does have if *A* were to interact with the same things (in the present case, with the same *mental* things) with which *B* does interact.

The box score is: the cross-context test shows that *certain differences that their effects have under intentional description* (namely, causing NaCl wants versus causing LCaN wants) and *certain differences that their effects have under nonintentional description* (see text) do *not* make the difference between having water thoughts and having twater thoughts a difference of causal power. We are about to see, however, that there are differences between the effects that twin mental states exhibit under intentional description which *survive* the cross-context test. These are the ones that, prima facie, make trouble for individualism.

[4] This, surely, is the intuitively natural way to compare causal powers. Consider:

—"Cats raised in Manhattan unable to climb trees," top scientist says.
—Why, that's rather disturbing. How do you explain it?
—There aren't any trees in Manhattan for them to climb.
—Oh.

(For more on this, see the Appendix)

So then, the question about individualism is: Do the twins' mental states belong to different natural kinds (do they have different causal powers) in virtue of differences in the intentional properties of the twins' behavior for which they are responsible? Since 'causal power', 'natural kind', and the like are, of course, technical terms, this question is not possessed of the highest degree of clarity. We shall see, however, that there are some clear intuitions about cases, and they will do for the purposes at hand.

It should be evident that the kind of question we are raising about the intentional states of twins can also arise in cases that have nothing in particular to do with intentionality. Suppose we have a pair of causes $C1$, $C2$, together with their respective effects $E1$, $E2$. Assume that:

> $C1$ differs from $C2$ in that $C1$ has cause property $CP1$ where $C2$ has cause property $CP2$.
>
> $E1$ differs from $E2$ in that $E1$ has effect property $EP1$ and $E2$ has effect property $EP2$.
>
> The difference between $C1$ and $C2$ is responsible for the difference between $E1$ and $E2$ in the sense that, if $C1$ had had $CP2$ rather than $CP1$, then $E1$ would have $EP2$ rather than $EP1$; and if $C2$ had had $CP1$ rather than $CP2$, $E2$ would have had $EP1$ rather than $EP2$.[5]

Call this *schema* S.[6] And now, what we want to know is: Which instances of schema S are cases where the difference between having $CP1$ and having $CP2$ is a difference in causal power in virtue of its responsibility for the difference between $E1$ and $E2$? (I shall often abbreviate this to 'When is having $CP1$ rather than $CP2$ a causal power?') If we knew the answer to this general question, then we would know whether, in particular, the difference between having water thoughts and having twater thoughts is a difference in causal power in virtue of its being responsible for the difference between my producing water behaviors and my twin's producing twater behaviors.[7] And, if we knew *that* we would know whether individualism is true; which is what we started out wanting to know, as the patient reader may recall.

---

[5] It is important, in order that relevant questions not be begged, that this is *all* that is required for the fact that $C1$ has $CP1$ to be "responsible for" the fact that $E1$ has $EP1$ (and similarly, mutatis mutandis, for the fact that $C2$ has $CP2$ to be responsible for the fact that $E2$ has $EP2$.) In like spirit, nothing in the examples will depend on stressing the requirement that the *C*s and the *E*s be related as causes and effects, so long as it is assumed that the difference between the *C*s is responsible for the difference between *E*s in the sense just specified.

[6] It will ease the exposition if we think of schema S sometimes as relating events and sometimes as relating event types. I shall exploit this ambiguity in what follows; but nothing in the argument turns on it.

Now, a plausible first reaction to this is that it is just not a philosophical issue. For, one might say, the question being raised is when the fact that a generalization supports counterfactuals shows that it has explanatory status. And the answer will depend, in the usual way, on systematic questions about the simplicity, plausibility, power, and so forth of the explanations, and on whether alternative counterfactual supporting generalizations are available. Some of the things that Tyler Burge[8] says about the desirability of not apriorizing about the taxonomy that psychological explanation requires suggest that he would approve of this line of thought; for example:

> It is a mistake . . . to allow ontological preconceptions that have only philosophical underpinning to affect one's interpretation of scientific enterprises. It is a larger mistake to allow them to dictate the sorts of explanatory kinds that are deemed admissible for explanation (*ibid.*).

But though I would not have thought that apriorism is among my major methodological vices, still this does strike me as perhaps a little prim. It would be surprising if we could give interesting a priori *sufficient* conditions for when a difference in the properties of causes constitutes a difference in their causal powers. But it seems likely that we can give a priori arguments for some *necessary* conditions. After all, the commitment to causal explanation presumably has *some* methodological consequences qua commitment to causal explanation; and it ought to be possible to tease these out by reflecting on what kinds of things causal explanations are.

And, in fact, it does seem reasonably clear, a priori, that some instances of schema *S* are not bona fide. Consider, for a thoroughly trivial example, the case where *EP*1 is the property of being the effect of *C*1, and *EP*2 is the property of being the effect of *C*2. There is, of course, a property of *C*1 in virtue of which its effects are effects of *C*1, namely, the property of being *C*1. Correspondingly, there is a property of *C*2 in virtue of which its effects are effects of *C*2, namely, the property of being *C*2. Counterfactual support goes through in the required way; if the cause of *E*1 had had the property of being *C*2 rather than *C*1, then *E*1 would have had the property *EP*2 rather than the property *EP*1. But it seems a priori obvious that this is not a case where having *CP*1 is a causal power of *C*'s in virtue of its responsibility for *E*'s having *EP*1. One of the properties of my effects that your effects cannot have, however hard you try, is *the property of being caused by me*. But I take it to be simply obvious that this difference in our effects does not make the property of *being me rather than you* a causal power. I am not a unit natural kind in virtue of my unique power to cause effects that are effects of me.

---

[7] Reminder: 'Water behavior' means not *behavior that has to do with water* but *behavior that has a reference to water in its intentional description*. We are assuming (see above) that the only relevant descriptions under which the behaviors of twins differ are intentional.

[8] "Individuation and Causation in Psychology" (manuscript, UCLA, 1989).

It seems clear a priori, then, that not every case in which a difference between causes is responsible for a difference in effects is a case where the difference in causes is a difference in their causal powers. There are, in fact, plenty of examples. I can define a property *being an H-particle* that is satisfied by any *x* at time *t* iff [(*x* is a physical particle at *t*) & (the coin in my hand at *t* is heads up)]. And correspondingly for T-particles. So a difference between properties of coins in my hand (namely, the difference between being heads up and being tails up) is responsible for the difference between the state of affairs in which all the particles in the universe are H-particles and the state of affairs in which all the particles in the universe are T-particles. But, of course, the difference between being heads up and being tails up does not count as a causal power in virtue of its responsibility for this difference in the particles.

Or again: in virtue of my having siblings, I am able to have sons who are nephews. A molecular twin who did not have siblings would ipso facto fail to have nephews among his children. But I take it to be a priori obvious that the difference between having siblings and not having siblings does not constitute a difference in the causal powers of parents in virtue of its responsibility for this difference among the properties of their offspring.[9]

While we are on the topic of what is a priori obvious: the preceding three examples might suggest that the difference between *CP*1 and *CP*2 in instances of schema *S* does not constitute a difference of causal powers when *CP*1 and *CP*2 are *relational* properties. But a moment's reflection shows that this cannot be right (as, indeed, *Psychosemantics* was at some pains to point out). Taxonomy by relational properties is ubiquitous in the sciences, and it is not in dispute that properties like *being a meteor* or *being a planet*—properties which could, notice, distinguish molecularly identical chunks of rock—constitute causal powers. It is because *this* rock-twin is a planet and *that* rock-twin is not that this rock-twin has a Keplerian orbit and that rock-twin does not; it is because this rock-twin is a meteor and that rock-twin is not that this rock-twin's effects include craters and that rock-twin's effects do not. But, patently, *being a planet* and *being a meteor* are relational properties in good standing. To be a planet is to be a rock (or whatever) that is revolving around a star; to be a meteor is to be a rock (or whatever) that is falling, or has fallen, into collision with another rock.

---

[9] Another a priori taxonomic intuition that cries out to be taken seriously; it is (as *Psychosemantics* urged) preposterous to suggest that neurological (or biochemical; or molecular) states should be taxonomized by reference to the sorts of properties that distinguish twins in the standard examples—by whether there is water or twater in the local puddles, for example. Burge remarks that perhaps all this shows is that psychological taxonomy is more contextually sensitive than neurological taxonomy. But surely that will not run: if someone were to discover (for example) that living near high-tension wires turns your dendrites green, you can bet the neurologists would pay attention. So: what is the difference between living near high-tension wires and living near puddles of twater such that one, but not the other, is a candidate for a neurological causal power? (See below.)

Since the intuitions are pretty strong in all these cases, there is prima facie reason to suppose that there are some conditions on cause properties being causal powers which can be recognized from the armchair: not all cause properties are causal powers; not all relational properties fail to be, for two examples. I shall presently state a condition that, I claim, has to be satisfied if a property of a cause is a causal power in virtue of its responsibility for a certain property of an effect; and I shall claim that properties that distinguish twins (like being causally connected to water rather than twater; or having water thoughts rather than twater thoughts) do not satisfy this condition in virtue of their responsibility for differences between the intentional properties of the twins' behaviors. My evidence for the acceptability of this condition will be largely that it sorts examples like the ones I have just run through in an intuitively satisfactory way.

We need, however, some more ground clearing before we get down to it. (I solicit the reader's forbearance; once the ground is all clear, we shall be able to move very fast.) To begin with, I want to call your attention to what turns out to be a critical property of schema *S*: the question we are raising is not whether the difference between having *CP*1 and having *CP*2 is a difference in causal powers; rather, it is whether the difference between having *CP*1 and having *CP*2 is a difference in causal powers *in virtue of its being responsible for a certain difference between E1 and E2*, namely, in virtue of its being responsible for *E*1's having *EP*1 rather than *EP*2 and for *E*2's having *EP*2 rather than *EP*1. The point I am wanting to emphasize is that a cause property might fail to count as a causal power in virtue of its responsibility for one effect property, but still might constitute a causal power in virtue of its responsibility for some other effect property.

My coin's being heads up (rather than tails up) does not constitute a causal power in virtue of its responsibility for the universe being populated with H-particles; but it might constitute a causal power in virtue of its being responsible for my coin's reflecting light the way it does rather than some other way (e.g., rather than the way it would if it were tails up). Similarly, my having siblings does not constitute a causal power in virtue of its enabling me to have sons who are nephews. But suppose there is such a thing as sibling's disease; it causes people who have siblings to break out in a rash. Having a sibling might then be a causal power in virtue of being responsible for people coming down with sibling's disease. I stress this point because, if you do not relativize the issue about causal powers in this way, it turns out on a weak assumption that every contingent property is a causal power. The weak assumption is that it is (nomologically) possible to build a detector for any contingent property.

Consider, for example, the property of having had a Bulgarian grandmother who once plucked a daffodil as she went trippingly on her way to market. This is the sort of property that can distinguish between you and your molecular twin. And, no doubt, your effects have properties because of your having had such a grandmother which his do not have because he did not, e.g., your effects have one and all got the property of being the effects of someone who had a Bulgarian grandmother who . . ., etc. But, intuitively, that is not sufficient for making the possession of a Bulgarian *G* who . . .—or the lack of a Bulgarian *G* who . . .—a causal power. Indeed, you would not have thought that *any* of the Bulgarian-grandmother-dependent properties of your effects—any of the properties that distinguish the effects of the molecular twin with the Bulgarian *G* . . . from the effects of the molecular twin without one—would be of the right kind to make having a Bulgarian grandmother a causal power.

And yet, it would surely be possible to build (more precisely, it would surely have been possible to have built) a machine which exhaustively examines the piece of space-time that starts with the birth of your grandmother and ends with your birth and which goes into one state if it detects somebody who was your grandmother and was Bulgarian and once plucked a daffodil as she went trippingly on her way to market, but which goes into a different state in case it detects no such person. I assume that it is possible (in principle) to build a machine that reliably detects this property. I am prepared to assume that it is possible (in principle) to build such a machine for any contingent property at all. If this assumption is true, then the Bulgarian grandmother detector can distinguish between you and your molecular twin, and having a Bulgarian grandmother is having a causal power *in virtue of the (actual or possible) effects that instantiations of this property have on Bulgarian G . . . detectors*. (Quite generally, if any contingent property can be detected then any contingent property is a causal power in virtue of the effects of its instantiations on its detectors.) What does not follow, however, is that having had a Bulgarian *G* . . . is a causal power *in virtue of its effects on your behavior;* or, indeed, in virtue of its responsibility for *any property of yours*.

I am going through this song and dance by way of replying to an argument that Burge suggested (but does *not* endorse) that might lead one to think that having a water thought (rather than a twater thought) *must* be a causal power. After all, this argument says, it must be possible, in principle, to build a machine which looks through the chunk of space-time that starts with my birth and ends up with now, and which goes into one state if I am connected (in the right way) to water but goes into another state if I am connected (in the right way) to twater. This machine responds differently to me and my molecular twin, and does so in virtue of the fact that I am connected to water in the way that my twin is connected to *XYZ*. So my twin and I differ in our power to effect the states of this machine. So having water thoughts rather than twater thoughts (namely, having the kind of thoughts that you have if you are connected to water rather than twater) is a causal power in virtue of its responsibility for the machine's being in the state it is. So, if individualism says that having water thoughts rather than twater thoughts is *not* a causal power, then individualism is false.

Individualism does *not* say, however, that having water thoughts rather than twater thoughts is not a causal power. What it says is that having water thoughts rather than twater thoughts is not a causal power *in virtue of its being responsible for your producing water behaviors rather than twater behaviors*. (And similarly, being connected to water rather than twater is not a causal power in virtue of its being responsible for your having water thoughts rather than twater thoughts.) This is to say that the difference between having water thoughts and having twater thoughts is not a causal power in virtue of its responsibility for those of your properties *which are relevant to what psychological natural kinds your thoughts belong to*, e.g., it is not a causal power in virtue of its responsibility for the properties of your behavior.

That is all that an individualist has to show; to ask him to show more would be to make individualism false if there can be detectors for broad content properties. But that would be to trivialize the issue about individualism[10] since, as remarked above, it is plausible that there can be detectors for *any* contingent property (and, up to the limits of Turing machines, for quite a lot of noncontingent properties, too.)

So now: two causes differ in a certain property, and their effects differ in a certain property in virtue of this difference in the causes, and we want to know when the difference between the effects makes the difference between the causes a causal power. I shall tell you in just a moment. But it may be worth emphasizing that the cases we are interested in are not ones where the property that distinguishes the causes is *itself* the property of having a certain causal power.

Consider the case where $CP1$ just is the property of having the causal power to produce events that have $EP1$. Then, of course, the difference between having $CP1$ and not having it is the difference between having a certain causal power and not having it. This sort of case is easy because *it is noncontingent* that having $CP1$ is having a causal power. There are other, slightly less transparent, cases of this kind. For example, it is noncontingent that *being soluble in water* is having a causal power, because it is noncontingent that things that have that property have the power of dissolving in water; and the property of *being a camshaft* is a causal power because it is noncontingent that things that have that property have the power to lift the valves in a certain kind of engine (given conditions of optimal functioning . . ., etc.). There are, of course, scientifically interesting questions about properties that are noncontingently causal powers; for example, there are interesting questions about the supervenience bases of these properties and about their proper analysis into microfunctions. But there are not any interesting questions about whether things that have these properties have causal powers in virtue of having them. That question answers itself.

Compare, however, the properties of being a planet, being a meteor, and the like. These are causal powers in virtue of, for example, their respective abilities to produce Keplerian orbits and craters.

---

[10] Even being an H-particle would be a causal power in virtue of the nomological possibility of building H-particle detectors. Particle detecting is notoriously expensive and tricky; but once you have got a particle detector, it is trivial to convert it to detect H-particles. An H-particle detector is a particle detector that is hooked up to any gadget that is sensitive to which face of my coin is up. A particle detector that is hooked up to a Bulgarian grandmother would do.

But, insofar as being a planet is a causal power in virtue of the ability of planets to produce Keplerian orbits, it is contingent that being a planet is a causal power (for it is contingent that planets have Keplerian orbits); and insofar as being a meteor is a causal power in virtue of the ability of meteors to make craters, it is contingent that being a meteor is a causal power (for it is contingent that meteors make craters).

Notice that the broad content cases are like the meteor and planet cases and unlike the dispositional and functional cases. It may be that being connected to water rather than twater (hence having water thoughts rather than twater thoughts) is having a causal power; but if it is, it is contingent that it is. The property of being connected to water is not *identical* to the property of having a certain causal power, though it may be that there are causal powers that one has if one is connected to water that one would not have if one were not.

I draw two morals. First (contrary to some suggestions of R. Van Gulick's[11]), the fact that having a functional property is ipso facto having a certain causal power throws no particular light, one way or the other, on the question whether having a broad-content property is having a causal power. This conclusion may seem paradoxical since, after all, it is supposed to be that psychological properties *are* functional; so, if functional properties are noncontingently causal powers, and if psychological properties are functional, how could water thoughts and twater thoughts fail to be causal powers?

The answer is that it is actually not in dispute whether water thoughts and twater thoughts are causal powers. On the contrary, *of course* they are: my water thoughts are causally responsible for my reaching for water, my twin's twater thoughts are causally responsible for his reaching for twater . . . , and so on. The question on which local supervenience—hence individualism—turns, however, is whether the difference between water thoughts and twater thoughts is a *difference* in causal powers. The anti-individualist says 'yes, it is, in virtue of the intentional difference between the behaviors that water thoughts and twater thoughts cause'. The individualist says 'no it isn't. Water thoughts and twater thoughts are the same causal powers, only they're instantiated in people with different causal histories'. The idea that mental states are functional roles resolves this dispute only on the question-begging assumption that water behaviors and twater behaviors are behaviors of different kinds.

---

[11] "Metaphysical Arguments for Internalism and Why They Don't Work," in S. Silvers, ed., *Rerepresentation* (Philosophical Studies Series, #40) (Boston: Kluwer, 1989).

Imagine that we had different words for *being thirsty and born in the Bronx* and *being thirsty and born in Queens*. There would then be no question but that being Bronx-thirsty and being Queens-thirsty are causal powers; being Bronx-thirsty and being Queens-thirsty both make people drink, for example. But a question might arise whether being Bronx-thirsty and being Queens-thirsty are *different* causal powers.[12] My point is that this issue would still be open *even though* there is no question but that being Bronx-thirsty and being Queens-thirsty are causal powers.

A functionalist might undertake to argue that they are different causal powers because, on the one hand, psychological states are functionally individuated, and, on the other hand, Bronx thirst and Queens thirst lead to different behavioral consequences, namely, to Bronx thirst-quenching behaviors in the one case and Queens thirst-quenching behaviors in the other. But, clearly, to argue this way would be to beg the question. For, anybody who denies that Queens thirst and Bronx thirst are different causal powers, will also deny—and for the same reason—that Bronx thirst-quenching behavior and Queens thirst-quenching behavior are behaviors of different kinds. So, the principle that psychological states are functionally individuated does not settle questions about identity and difference of causal powers in this sort of case. And, of course, for an individualist, 'wants water' *is* this sort of case; for an individualist, 'wants water' means something like "thirsty and born *here*."

What distinguishes twins and threatens local supervenience is the property of *having water thoughts rather than twater thoughts*. No doubt, if functionalism is true, then mental states are one and all causal powers. But of course it does not follow from functionalism that differences between mental states are one and all functional differences; so it does not follow that differences between mental states are one and all differences of causal power. In fact, so far as I can tell, the only relevant connection between functionalism and the present issues about individualism is this: if, as I have claimed, the difference between the mental states of twins is at best *contingently* a difference of causal powers, it follows that it *cannot* be a difference of functional role, since differences of functional role are differences of causal powers *non*contingently.[13] Compare being water soluble rather than twater soluble. It is *not* contingent that having *this* property is having a causal power; to be soluble in water but not in twater just is to have the power to dissolve in the first but not in the second. So it is all right to allow that *being water soluble rather than twater soluble* is a functional property, as, indeed, intuition demands.

---

[12] This might well turn out to be the same question as whether 'is Bronx-thirsty' and 'is Queens-thirsty' are projectible; whether, for example, psychological generalizations about Bronx-thirsty people as such are confirmed by their instances.

[13] This comports with the idea that, whereas *having a belief* (desire, whatever) is being in the right functional state, having a belief *that P* is a matter of having the right head-to-world relations. So the difference between having a belief that P and a belief that Q *is not* a functional difference. For more on this, see *Psychosemantics*.

My impression is that this is all reasonably untendentious. In general, friends of broad content argue that it perfectly well *could turn out*—that there is no metaphysical reason why it should not turn out—that having mental states that differ in their broad content is having mental states that differ in their causal powers. (For example, it could turn out that there are causal laws that distinguish between twins.) But I do not remember hearing anyone argue that having mental states that differ in the way that the mental states of twins do *just is* having mental states that differ in their causal powers. On the contrary, according to the usual understanding, it is their causal *histories* that distinguish the mental states of twins. And the intuition about features of causal history is that some of them are causal powers (e.g., *having been dropped in transit; having been inoculated for smallpox*) and some of them are not (e.g., *having had a Bulgarian grandmother; having been born on a Tuesday*) and it is contingent which are which.

So, then, the second moral that I draw is that, in our hunt for a useful condition on what makes a difference in causal properties a difference in causal powers, we can restrict ourselves to cases where its *contingent* whether the difference in the properties constitutes a difference between powers. We are, finally, getting to where you can smell blood.

I am now, at last, going to tell you a story about why being a planet (for example) is a causal power and having siblings (for example) is not. I propose to do this in two steps. First, I shall tell you a simplified version of the story; it has the virtue of making the basic idea relatively transparent, but it has the disadvantage that it does not work. I shall then make the technical moves required to plug the leak; the complicated version that emerges will be a motivated condition upon a property of a cause being a causal power; one which, I claim, broad-content properties fail to meet.[14]

---

[14] Perhaps I had best reiterate that this is short for: ". . . a motivated condition for a difference between properties of causes being a difference in their causal powers" and that the claim is not that being a water (/twater) thought is not a causal power, but rather that the difference between having a water thought and having a twater thought is not a difference in causal power. More generally, the claim is that no two states differ in their causal powers *just* in virtue of differing in their broad contents.

So, then: here is me and here is my molecular twin; and I have siblings and he does not; and in virtue of my having siblings my sons are nephews, and in virtue of his not having siblings his sons are not nephews; and what we want to know is: Why is *having siblings* not a causal power in virtue of its being responsible for this difference in our offspring? Here is a first fling at the answer: it is because having siblings is *conceptually* connected to having sons who are nephews; to be a nephew *just is* to be a son whose parents have siblings. And, to put it roughly, your causal powers are a function of your *contingent* connections, not of your conceptual connections. As, indeed, Uncle Hume taught us.

Similarly: though it is a fact that all the world's particles become H-particles when my coin is heads up, that fact does not make being heads up a causal power of my coin. That is again because the connection between all the world's particles becoming H-particles at *t* and my coin's being heads up at *t* is conceptual. To be an H-particle at *t just is* to be a particle at a time when my coin is heads up.

Compare cases of relational properties that really are causal powers, like *being a planet*. Being a planet is a causal power in virtue of, for example, its contingent (a fortiori, nonconceptual) connection with having a Keplerian orbit. That is, being a planet is a causal power because it is true and contingent that, if you have molecularly identical chunks of rock, one of which is a planet and the other of which is not, then, ceteris paribus, the one which is a planet will have a Keplerian orbit, and ceteris paribus, the one which is not a planet will not.

Here is the general form of the proposed solution.[15] Consider an instance of schema *S*. *C*1 has *CP*1, *C*2 has *CP*2, *E*1 has *EP*1, *E*2 has *EP*2, and the difference between the causes is responsible for the difference between the effects in the sense that *E*1 would not have had *EP*1 (rather than *EP*2) but that *C*1 had *CP*1 (rather than *CP*2). And what we want to know is: When does the fact that this difference in the causes is responsible for this difference in the effects make *CP*1 and *CP*2 causal powers? The answer, which I shall call *condition C*, is:

> Only when it is not a conceptual truth that causes that which differ in that one has CP1 where the other has CP2 have effects that differ in that one has EP1 where the other has EP2.[16]

---

[15] But please, *please* keep it in mind that this solution applies in cases where, if the properties under consideration are causal powers, then they are causal powers contingently. In particular, it is *not* supposed to apply to properties that are causal powers necessarily, like dispositions. See above.

[16] I want to emphasize that condition *C* does not impugn the Davidsonian doctrine that the necessity/contingency of relations among *events* is description-rela-

So, for example, it is all right with condition C for *being a meteor* to be a causal power in virtue of the fact that meteors are responsible for craters. Take a pair of rock-twins such that one is a meteor and the other is not. Then (ceteris paribus) craters will be among the effects of the first but not among the effects of the second. And the relation between the difference between the rock-twins and the difference between their effects is nonconceptual. So it is all right for being a meteor (rather than a meteor twin) to be a causal power in virtue of the fact that meteors cause craters and meteor twins do not.[17]

Notice that the moral is that it is all right for being a meteor to be a causal power in virtue of this fact; not that being a meteor *is* a causal power in virtue of this fact. Satisfying condition C is, I suppose, necessary but not sufficient for being a causal power. As we are about to see, however, broad-content properties *fail* to satisfy this necessary condition, and that is enough to vindicate individualism.

Consider, first, the property of having water in your history (the property of being connected to water in whatever way it is that I am and my twin is not). The difference between being so connected and not being so connected is responsible for a certain difference between the broad contents of my thoughts and the broad contents of my twin's, namely, that I have water thoughts and he has twater thoughts. What we want to know is: Does this difference between our histories count as a causal power in virtue of the difference between the contents of our thoughts for which it is responsible. And the answer is: "No, because it is *conceptually necessary* that if you are connected to water in the right way then you have water thoughts (rather than twater thoughts) and it is again conceptually necessary that if you are connected to twater in the right way then you have

---

tive. Suppose $e1$ causes $e2$, so that '$e1$ causes $e2$' is contingently true. Still, there will be descriptions satisfied by $e1$ which entail (/presuppose) that $e1$ causes $e2$ (for example, such descriptions as 'the cause of $e2$'; 'the cause of $e2$ caused $e2$' is, of course, necessarily true). But none of this implies that there is anything description-relative about whether the instantiation of one property entails the instantiation of another (about whether, for example, it is conceptually necessary that whatever instantiates bachelorhood instantiates unmarriedness; or whether it is conceptually necessary that whatever instantiates *water-thinker-hood* instantiates *causally-connected-to-water-hood*. It is these latter sort of claims, not the former sort, which are at issue in respect of condition C.

[17] By contrast, I take it that their ability to produce *meteor* craters (where a meteor crater just is a crater that is caused by a meteor) is not a causal power that meteors have over and above their power to produce craters *tout court*. This is for the now familiar sort of reason: it is conceptually necessary that, whereas a crater caused (in the right way) by a meteor is a meteor crater, a crater caused (in whatever way) by a molecularly identical nonmeteor is not. Analogously, its ability to produce sunburns is not a causal power that the sun has over and above its power to produce burns.

110146

twater thoughts (rather than water thoughts)." To have a water thought *just is* to have a thought that is connected to water in the right way, and to have a twater thought *just is* to have a thought that is connected to twater in the right way.[18] So it is not the case that my being connected to water rather than to twater is a difference in my causal powers in virtue of its responsibility for my having water thoughts rather than twater thoughts.

Now consider the difference between having water thoughts and having twater thoughts. Water thoughts cause water behavior (drilling for water and the like); twater thoughts cause twater behavior (drilling for twater and the like). What we want to know is: Does the difference between having water thoughts and having twater thoughts count as a causal power in virtue of the fact that it is responsible for this difference in the intentional properties of the behavior of the thinker? And the answer is: "No, because it is conceptually necessary that people who have water thoughts (rather than twater thoughts) produce water behavior (rather than twater behavior)." Being water behavior (rather than twater behavior) *just is* being behavior that is caused by water thoughts (rather than twater thoughts). So, though it is true that water thoughts are responsible for water behavior and twater thoughts are not, it does not follow that water thoughts have a causal power that twater thoughts lack. On the contrary, being a water thinker is the same causal power as being a twater thinker, only instantiated in a person with a different causal history.

I have been saying that it is only when the difference between the causes is not conceptually connected to the corresponding differences in the effects that the difference in the causes counts as a difference in causal powers. I think this really is the heart of the matter; it is why being responsible for particles being H-particles, being responsible for sons' being nephews, being responsible for behavior's being water behavior, and the like do not count as causal powers. But alas, the proposal does not work as stated and I shall have to do some patching.

Suppose that water is Bush's favorite thing to drink. Then, we have both:

1. If I am connected to water in the right way then my thoughts are water thoughts.

and

---

[18] More precisely (and assuming that functionalism is right about what beliefs, desires, and the like are) to have a water thought is to have a thought that (a) is connected to water in the right way; and (b) has whatever functional properties water thoughts and twater thoughts share. I leave out (b) in the text to simplify the exposition. No doubt the exposition could do with some of that.

2. If I am connected to water in the right way, then my thoughts are thoughts about Bush's favorite drink.

Notice that both 1 and 2 distinguish me from my twin: because he is not connected to water in the right way, it is false of him that he has water thoughts, and it is also false of him that his thoughts are about *B*'s favorite drink. And, though 1 is conceptually necessary, 2 is contingent. That is, even though the difference between being water-connected and being twater-connected does not satisfy condition *C* in virtue of its being responsible for the difference between having water thoughts and having twater thoughts, it *does* satisfy condition *C* in virtue of its responsibility for the difference between having thoughts that are about Bush's favorite drink and having thoughts that are not. Hence, it satisfies condition *C* in virtue of making some prima facie *psychological* difference, some difference that is prima facie relevant to psychological taxonomy. So, even if, as I claim, condition *C* has its heart in the right place, still, as stated it has no teeth. Damn![19]

Everything is going to be all right, however. Let *G* be a property that nephews have, and let it be as contingent as you like that nephews have it. (*G* might be the property of glowing in the dark if it turns out that nephews do that.) Then the following is necessary:

3. If *G* is a property that nephews have, then if I have siblings then my sons have *G*.

So, for example, it is necessary that (if nephews glow in the dark, then if I have siblings, then my sons glow in the dark). The reason that this is necessary is, roughly, that it is conceptually necessary that the sons of people with siblings are nephews, so it is conceptually necessary that if you have a sibling then your sons have whatever properties nephews have. Notice, to repeat, that this is true even of properties that nephews have contingently: it is conceptually necessary that if *P* is a property that nephews have contingently then if you have siblings then your sons have *P*. (What *is not* true, of course, is 3′, the variant of 3 that has the modal operator imported. 3′ is false in the cases where *P* is a property that nephews have contingently.)

3′. If *P* is a property of nephews, then if you have siblings then it is necessary that your sons have *P*.

---

[19] I am indebted to Stich for this line of argument. I suppose he thinks I should be grateful.

Now compare 4, which I take to be clearly contingent.

4. If *E* is a property that Keplerian orbits have then, if I am a planet, then my orbit has *E*.

(You might wish to try out 4 reading *E* as the property of being eliptical.)

The reason that 4 is contingent is, roughly, that it is contingent that planets have Keplerian orbits, so it is contingent that if I am a planet then my orbit has whatever properties Keplerian orbits do. This is true even if *E* is a property that Keplerian orbits have necessarily (like being a Keplerian orbit).[20]

Now we can do the case that was bothering us before. Consider 5, where *B* might be the property of being a thought about Bush's favorite drink:

5. If *B* is a property that water thoughts have, then if I am connected to water in the right way, then *B* is a property that my thoughts have.

I take it that 5 is obviously conceptually necessary. The reason is, roughly, that it is conceptually necessary that if I am connected to water in the right way then my thoughts are water thoughts, and it is a truism that if something is a property of water thoughts then it is a property of my thoughts if my thoughts *are* water thoughts.[21] Similarly, mutatis mutandis, with 6.

6. If *B* is a property that water behaviors have, then if my thoughts are water thoughts then my behaviors have *B*.

6 is conceptually necessary because, roughly, it is conceptually necessary that the behaviors that water thoughts cause are water behaviors, so it is conceptually necessary that if something is a property of water behaviors then it is a property of the behaviors of water thinkers. This is true even if *B* is a property that water behaviors have contingently, like being behaviors that are concerned with Bush's favorite drink.

Compare these cases with ones where thoughts have bona fide causal powers. Suppose that thinking about topology gives one headaches. Then we have 7:

---

[20] 4 is necessary, however, if you choose *E* as some property that *everything* has necessarily; like being self-identical. This does not, of course, prejudice the present point.

[21] For those following the technicalities: in effect, my original line of argument depended on there being a conceptual relation between water connectedness and water thoughts construed *de dicto*. Stich gets a *contingent* relation between water connectedness and water thoughts *de re* by invoking such contingent premises as, for example, that water is Bush's favorite drink; and thus trivially satisfies condition *C*. In effect, 5 provides a substantive version of *C* by conditionalizing on these contingent premises.

7. If *B* is a property of headaches then if I have topology thoughts then my mental state has *B*.

7 is clearly contingent, and it is so because of the contingency of the (putative) relation between something's being a topology thought and its having headaches among its effects. Notice that 7 is contingent even if *B* is a property that headaches have necessarily (like being headaches).

So, here is the story. For the difference between being *CP*1 and being *CP*2 to be a difference of causal powers, it must at least be that the effects of being *CP*1 differ from the effects of being *CP*2. But, I claim, it is further required that this difference between the effects be *nonconceptually* related to the difference between the causes. This further condition is motivated both by our intuitions about the examples and by the Humean consideration that causal powers are, after all, powers to enter into nonconceptual relations. Broad-content differences, per se, do not satisfy this condition, however. There are differences between my behavior and my twin's which are due, in the first instance, to the difference between the intentional contents of our thoughts, and, in the second instance, to my being connected to water in a way that he is not. But these differences among the effects are conceptually related to the differences between the causes; it is conceptually necessary that being connected to water rather than twater leads to water thinking rather than twater thinking; and it is again conceptually necessary that water thinking leads to water behaving and twater thinking does not.

So, then, the difference between the mental states of twins does not count as a difference in causal power in virtue of its responsibility for the intentional differences among twin behaviors. So argument *B* is no good; what is wrong with it is that the inference from 2′ to 3′ is unsound.[22] Finally, since it is assumed that the effects of mental states that differ only in broad content are (relevantly) different *only*

---

[22] I can imagine that someone might now want to say: "All this follows from my conceding that the intentional properties that distinguish twins are causal powers contingently if they're causal powers at all. Well, I was wrong; those sorts of differences among broad content states are *non*contingently differences of causal powers. For example, the property of having water thoughts (rather than twater thoughts) is *identical* to the property of being able to produce water behaviors (rather than twater behaviors); the property of being connected to water (rather than twater) is *identical* to the property of being able to have water thoughts (rather than twater thoughts), etc. Notice that being soluble is a causal power even though the difference between being soluble and not is conceptually connected to the difference between dissolving and not."

All right, but it really does not help; if having water thoughts (rather than twater thoughts) is identical to having (inter alia) the power to drill for water (rather than twater), then the broad-content psychological generalizations that distinguish twins under intentional description, it follows that there are *no* taxonomically relevant differences consequent upon broad-content differences as such. From the point of view of psychological taxonomy, my mental states must therefore belong to the same natural kind as those of my molecular twin. So individualism is true and local supervenience is preserved. End of story.

—Oi! Halt thief! Stop that person.
—What's the matter?
—You promised me an argument *for* narrow content. But all you've given me is an argument *against* arguments against individualism. I want my money back.
—Sorry.

We have seen that twater thoughts and water thoughts are not different causal powers. So, for the psychologist's purposes, they are the same intentional state.[23] But they cannot be the same intentional state unless they have the same intentional content. And they cannot have the same intentional content unless intentional content is individuated narrowly. Now it is an argument for narrow content.

—Much obliged.
—My pleasure.

JERRY A. FODOR

Rutgers University and CUNY Graduate Center

APPENDIX: A NOTE ON THE CROSS-CONTEXT TEST FOR IDENTITY OF CAUSAL POWERS

Suppose that my twin and I are both green, but I live in red world (where everything except me is red) and he lives in green world (where everything including him is green). So then, he has a property of *being adapted to the*

[like, 'if you have water wants (rather than twater wants) then you drill for water (rather than twater)'] themselves all come out conceptually necessary (a kind of point to which Ryleans were, of course, entirely alert). So, if the assumption is that intentional states are noncontingently causal powers, then the appropriate form for the individualist's supervenience claim is that no *contingent* intentional generalization can distinguish twins (no contingent intentional generalization can be such that one but not the other of the twins satisfies its antecedent or consequent). The moral would then be: no causal laws about broad intentional states as such (in fact, just the sort of moral that Ryleans used to draw).

It is true that being soluble is a causal power even though it is conceptually connected to dissolving. But the price for thus evading condition *C* is the "quasi-logical" status of 'if soluble then dissolves'.

[23] —But why can this putative psychologist not allow them to be *different* states but with the *same* causal powers?

—Because, if he does, his theory misses generalizations, namely, all the generalizations that subsume me and my twin. Good taxonomy is about *not* missing generalizations.

color of his environment which I lack. And, intuitively, this property is a causal power: because he has it, he can sneak up on things; because I lack it, I cannot. But this appears to make trouble for the idea that identity of causal powers has to be assessed *across* contexts. The cross-context test implies that there cannot be a difference in our causal powers unless there are counterfactuals that are true of one of us but not of the other; but, in the present case, it looks like the counterfactuals come out the same for him as they do for me. If he were in my world, he could not sneak up on things; if I were in his world, I could. So our causal powers are the same according to the cross-context test. So there must be something wrong with the idea that identity of causal powers should be assessed across contexts. (For elaboration of what I take to be essentially this line of argument, see Van Gulick, *op. cit.*)

But this argument is fallacious; it rests on a confusion between, to put it roughly, an attributive and a referential way of reading the definite description in 'the color of his environment', with a consequent confusion about how the across-context test applies. This will only take a minute to sort out. If you read 'the color of his environment' referentially, then my twin's being adapted to the color of his environment is his being green. Read that way, then, *I too am adapted to the color of his environment* (I am green, too); so being adapted to the color of his environment is a causal power that we *share*. As, indeed, the cross-context test for causal powers properly predicts: if I were in his environment, I would be able to sneak up on things; if he were in mine, he would not. It is just that, since I am *not* in his environment and he is, being adapted to the color of his environment does him a lot more good than it does me.

If, however, you read 'the color of his environment' attributively, then his being adapted to the color of his environment consists in his being the *same color as the environment that he is in*. So, if you move *that* property across contexts, then when he is in *my* environment his being adapted to the color of his environment consists in his being (not green like me but) *red*. So, if he is adapted to the color of his environment in *that* sense, then when we are both in my environment, he can sneak up on things and I cannot. So, according to the across-context test, being adapted to the color of his environment is a causal power that my twin has and I lack, which is, again, the intuitively correct result.

So there is nothing wrong with the test.

(—Well, but what about the property of *being-adapted-to-the-color-of-red-world-and-in-red-world*. Is *that* not a causal power that defies the cross-context test?

—Answer: It does not *defy* the cross-context test; it merely satisfies it trivially. Red-world is the only context in which one *can* have that property.)

REAL PATTERNS*

ARE there really beliefs? Or are we learning (from neuroscience and psychology, presumably) that, strictly speaking, beliefs are figments of our imagination, items in a superseded ontology? Philosophers generally regard such ontological questions as admitting just two possible answers: either beliefs exist or they do not. There is no such state as quasi existence; there are no stable doctrines of semirealism. Beliefs must either be vindicated along with the viruses or banished along with the banshees. A bracing conviction prevails, then, to the effect that when it comes to beliefs (and other mental items) one must be either a realist or an eliminative materialist.

I. REALISM ABOUT BELIEFS

This conviction prevails in spite of my best efforts over the years to undermine it with various analogies: are *voices* in your ontology?[1] Are *centers of gravity* in your ontology?[2]

It is amusing to note that my analogizing beliefs to centers of gravity has been attacked from both sides of the ontological dichotomy, by philosophers who think it is simply obvious that centers of gravity are useful fictions, and by philosophers who think it is simply obvious that centers of gravity are perfectly real:

> The trouble with these supposed parallels . . . is that they are all strictly speaking *false*, although they are no doubt useful simplifications for many purposes. It is false, for example, that the gravitational attraction between the Earth and the Moon involves two point masses; but it is a good enough first approximation for many calculations. However, this is not at all what Dennett really wants to say about intentional states. For he insists that to adopt the intentional stance and interpret an agent as acting on certain beliefs and desires is to discern a pattern in his actions which is genuinely there (a pattern which is missed if we instead adopt a scientific stance): Dennett certainly does not hold that the role of intentional ascriptions is merely to give us a useful approximation to a truth that can be more accurately expressed in non-intentional terms.[3]

* Thanks to Kathleen Akins, Akeel Bilgrami, Donald Davidson, Barbara Hannan, Douglas Hofstadter, Norton Nelkin, W. V. O. Quine, Richard Rorty, George Smith, Peter Suber, Stephen White, and the MIT/Tufts philosophy of psychology discussion group for the discussions that provoked and shaped this paper.

[1] *Content and Consciousness* (Boston: Routledge & Kegan Paul, 1969), ch. 1.

[2] "Three Kinds of Intentional Psychology," in R. Healey, ed., *Reduction, Time and Reality* (New York: Cambridge, 1981); and *The Intentional Stance* (Cambridge: MIT, 1987).

[3] Peter Smith, "Wit and Chutzpah," review of *The Intentional Stance* and Jerry A. Fodor's *Psychosemantics, Times Higher Education Supplement* (August 7, 1988), p. 22.

0022-362X/91/8801/27–51

Compare this with Fred Dretske's[4] equally confident assertion of realism:

> I am a realist about centers of gravity . . . . The earth obviously exerts a gravitational attraction on *all* parts of the moon—not just its center of gravity. The *resultant* force, a vector sum, acts through a point, but this is something quite different. One should be very clear about what centers of gravity are *before* deciding whether to be literal about them, *before* deciding whether or not to be a center-of-gravity realist (*ibid.*, p. 511).

Dretske's advice is well-taken. What are centers of gravity? They are mathematical points—abstract objects or what Hans Reichenbach called *abstracta*—definable in terms of physical forces and other properties. The question of whether abstract objects are real—the question of whether or not "one should be a realist about them"—can take two different paths, which we might call the metaphysical and the scientific. The metaphysical path simply concerns the reality or existence of abstract objects generally, and does not distinguish them in terms of their scientific utility. Consider, for instance, the *center of population* of the United States. I define this as the mathematical point at the intersection of the two lines such that there are as many inhabitants north as south of the latitude, and as many inhabitants east as west of the longitude. This point is (or can be) just as precisely defined as the center of gravity or center of mass of an object. (Since these median strips might turn out to be wide, take the midline of each strip as the line; count as inhabitants all those within the territorial waters and up to twenty miles in altitude—orbiting astronauts do not count—and take each inhabitant's navel to be the determining point, etc.) I do not know the center of population's current geographic location, but I am quite sure it is west of where it was ten years ago. It jiggles around constantly, as people move about, taking rides on planes, trains, and automobiles, etc. I doubt that this abstract object is of any value at all in any scientific theory, but just in case it is, here is an even more trivial abstract object: Dennett's lost sock center: the point defined as the center of the smallest sphere that can be inscribed around all the socks I have ever lost in my life.

These abstract objects have the same metaphysical status as centers of gravity. Is Dretske a realist about them all? Should we be? I do not intend to pursue this question, for I suspect that Dretske is—and we should be—more interested in the scientific path to realism: centers of gravity are real because they are (somehow) *good* abstract objects. They deserve to be taken seriously, learned about, used. If we go so far as to distinguish them as *real* (contrasting them, perhaps, with those abstract objects which are *bogus*), that is because we think they serve in perspicuous representations of real forces, "natural" properties, and the like. This path brings us closer, in any case, to the issues running in the debates about the reality of beliefs.

---

[4] "The Stance Stance," commentary on *The Intentional Stance*, in *Behavioral and Brain Sciences*, XI (1988): 511–2.

I have claimed that beliefs are best considered to be abstract objects rather like centers of gravity. Smith considers centers of gravity to be useful fictions while Dretske considers them to be useful (and hence?) real abstractions, and each takes his view to constitute a criticism of my position. The optimistic assessment of these opposite criticisms is that they cancel each other out; my analogy must have hit the nail on the head. The pessimistic assessment is that more needs to be said to convince philosophers that a mild and intermediate sort of realism is a positively attractive position, and not just the desperate dodge of ontological responsibility it has sometimes been taken to be. I have just such a case to present, a generalization and extension of my earlier attempts, via the concept of a *pattern*. My aim on this occasion is not so much to prove that my intermediate doctrine about the reality of psychological states is right, but just that it is quite possibly right, because a parallel doctrine is demonstrably right about some simpler cases.

We use folk psychology—interpretation of each other as believers, wanters, intenders, and the like—to predict what people will do next. Prediction is not the only thing we care about, of course. Folk psychology helps us understand and empathize with others, organize our memories, interpret our emotions, and flavor our vision in a thousand ways, but at the heart of all these is the enormous predictive leverage of folk psychology. Without its predictive power, we could have no interpersonal projects or relations at all; human activity would be just so much Brownian motion; we would be baffling ciphers to each other and to ourselves—we could not even conceptualize our own flailings. In what follows, I shall concentrate always on folk-psychological prediction, not because I make the mistake of ignoring all the other interests we have in people aside from making bets on what they will do next, but because I claim that our power to *interpret* the actions of others depends on our power—seldom explicitly exercised—to predict them.[5]

---

[5] R. A. Sharpe, in "Dennett's Journey Towards Panpsychism," *Inquiry*, XXXII (1989): 233–240, takes me to task on this point, using examples from Proust to drive home the point that "Proust draws our attention to possible lives and these

Where utter patternlessness or randomness prevails, nothing is predictable. The success of folk-psychological prediction, like the success of any prediction, depends on there being some order or pattern in the world to exploit. Exactly where in the world does this pattern exist? What is the pattern a pattern *of*?[6] Some have thought, with Fodor, that the pattern of belief must in the end be a pattern of structures in the brain, formulae written in the language of thought. Where else could it be? Gibsonians might say the pattern is "in the light"—and Quinians (such as Donald Davidson and I) could almost agree: the pattern is discernible in agents' (observable) behavior when we subject it to "radical interpretation" (Davidson) "from the intentional stance" (Dennett).

When are the elements of a pattern real and not merely apparent? Answering this question will help us resolve the misconceptions that have led to the proliferation of "ontological positions" about beliefs, the different grades or kinds of realism. I shall concentrate on five salient exemplars arrayed in the space of possibilities: Fodor's industrial-strength Realism (he writes it with a capital 'R'); Davidson's regular strength realism; my mild realism; Richard Rorty's milder-than-mild irrealism, according to which the pattern is *only* in the eyes of the beholders, and Paul Churchland's eliminative materialism, which denies the reality of beliefs altogether.

In what follows, I shall assume that these disagreements all take place within an arena of common acceptance of what Arthur Fine[7] calls NOA, the natural ontological attitude. That is, I take the interest in these disagreements to lie not in differences of opinion about the ultimate metaphysical status of physical things or abstract things (e.g., electrons or centers of gravity), but in differences of opinion about whether beliefs and other mental states are, shall we say, *as real as* electrons or centers of gravity. I want to show that mild realism is the doctrine that makes the most sense when what we are

---

possible lives are various. But in none of them is prediction of paramount importance" (240). I agree. I also agree that what makes people interesting (in novels and in real life) is precisely their unpredictability. But that unpredictability is only interesting against the backdrop of routine predictability on which all interpretation depends. As I note in *The Intentional Stance* (p. 79) in response to a similar objection of Fodor's, the same is true of chess: the game is interesting only because of the unpredictability of one's opponent, but that is to say: the intentional stance can usually eliminate *only* ninety percent of the legal moves.

[6] Norton Nelkin, "Patterns," forthcoming.

[7] *The Shaky Game: Einstein Realism and the Quantum Theory* (Chicago: University Press, 1986); see esp. p. 153n, and his comments there on Rorty, which I take to be consonant with mine here.

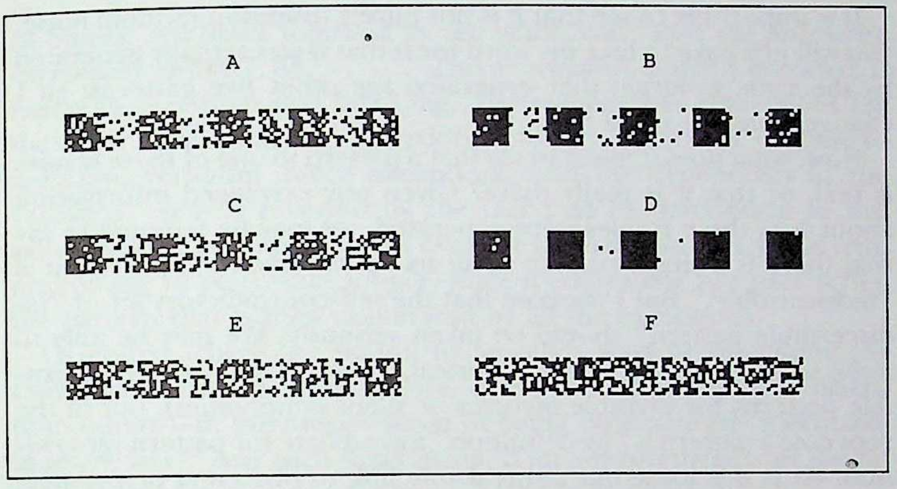

Figure 1

talking about is real patterns, such as the real patterns discernible from the intentional stance.[8]

In order to make clear the attractions and difficulties of these different positions about patterns, I shall apply them first to a much simpler, more readily visualized, and uncontroversial sort of pattern.

II. THE REALITY OF PATTERNS

Consider the six objects in Figure 1 (which I shall call *frames*):

We can understand a frame to be a finite subset of data, a window on an indefinitely larger world of further data. In one sense *A–F* all display different patterns; if you look closely you will see that no two frames are exactly alike ("atom-for-atom replicas," if you like). In another sense, *A–F* all display the same pattern; they were all made by the same basic process, a printing of ten rows of ninety dots, ten black dots followed by ten white dots, etc. The overall effect is to create five equally spaced black squares or bars in the window. I take it that this pattern, which I shall dub *bar code,* is a real pattern if anything is. But some random (actually pseudo-random) "noise" has been allowed to interfere with the actual printing. The noise ratio is as follows:

A: 25% *B*: 10%

C: 25% *D*: 1%

E: 33% *F*: 50%

---

[8] See *The Intentional Stance*, pp. 38–42, "Real patterns, deeper facts, and empty questions."

It is impossible to see that *F* is not purely (pseudo-) random noise; you will just have to take my word for it that it was actually generated by the same program that generated the other five patterns; all I changed was the noise ratio.

Now, what does it mean to say that a pattern in one of these frames is real, or that it is really there? Given our privileged information about how these frames were generated, we may be tempted to say that there is a single pattern in all six cases—even in *F*, where it is "indiscernible." But I propose that the self-contradictory air of "indiscernible pattern" should be taken seriously. We may be able to make some extended, or metaphorical, sense of the idea of indiscernible patterns (or invisible pictures or silent symphonies), but in the root case a pattern is "by definition" a candidate for pattern *recognition*. (It is this loose but unbreakable link to observers or perspectives, of course, that makes "pattern" an attractive term to someone perched between instrumentalism and industrial-strength realism.)

Fortunately, there is a standard way of making these intuitions about the discernibility-in-principle of patterns precise. Consider the task of transmitting information about one of the frames from one place to another. How many bits of information will it take to transmit each frame? The least efficient method is simply to send the "bit map," which identifies each dot *seriatim* ("dot one is black, dot two is white, dot three is white, . . ."). For a black-and-white frame of 900 dots (or pixels, as they are called), the transmission requires 900 bits. Sending the bit map is in effect verbatim quotation, accurate but inefficient. Its most important virtue is that it is equally capable of transmitting any pattern or any particular instance of utter patternlessness.

Gregory Chaitin's[9] valuable definition of mathematical randomness invokes this idea. A series (of dots or numbers or whatever) is random if and only if the information required to describe (transmit) the series accurately is *incompressible:* nothing shorter than the verbatim bit map will preserve the series. Then a series is not random —has a pattern—if and only if there is some more efficient way of describing it.[10] Frame *D*, for instance, can be described as "ten rows of ninety: ten black followed by ten white, etc., *with the following exceptions:* dots 57, 88, ..." This expression, suitably encoded, is much shorter than 900 bits long. The comparable expressions for the other frames will be proportionally longer, since they will have to mention, verbatim, more exceptions, and the degeneracy of the "pattern" in *F* is revealed by the fact that its description in this system will be no improvement over the bit map—in fact, it will tend on average to be trivially longer, since it takes some bits to describe the pattern that is then obliterated by all the exceptions.

---

[9] "Randomness and Mathematical Proof," *Scientific American*, CCXXXII (1975): 47–52.

[10] More precisely: "A series of numbers is random if the smallest algorithm capable of specifying it to a computer has about the same number of bits of information as the series itself" (Chaitin, p. 48). This is what explains the fact that the "random number generator" built into most computers is not really properly named, since it is some function describable in a few bits (a little subroutine that is called for some output whenever a program requires a "random" number or series). If I send you the description of the pseudo-random number generator on

Of course, there are bound to be other ways of describing the evident patterns in these frames, and some will be more efficient than others—in the precise sense of being systematically specifiable in fewer bits.[11] Any such description, if an improvement over the bit map, is the description of a real pattern in the data.[12]

Consider bar code, the particular pattern seen in *A–E,* and almost perfectly instantiated in *D. That* pattern is quite readily discernible to the naked human eye in these presentations of the data, because of the particular pattern-recognition machinery hard-wired in our visual systems—edge detectors, luminance detectors, and the like. But the very same data (the very same streams of bits) presented in some other format might well yield no hint of pattern to us, especially in the cases where bar code is contaminated by salt and pepper, as in frames *A* through *C*. For instance, if we broke the 900-bit series of frame *B* into 4-bit chunks, and then translated each of these into hexadecimal notation, one would be hard pressed indeed to tell the resulting series of hexadecimal digits from a random series, since the hexadecimal chunking would be seriously out of phase with the deci-

---

my computer, you can use it to generate exactly the same infinite series of random-seeming digits.

[11] Such schemes for efficient description, called compression algorithms, are widely used in computer graphics for saving storage space. They break the screen into uniformly colored regions, for instance, and specify region boundaries (rather like the "paint by numbers" line drawings sold in craft shops). The more complicated the picture on the screen, the longer the compressed description will be; in the worst case (a picture of confetti randomly sprinkled over the screen) the compression algorithm will be stumped, and can do no better than a verbatim bit map.

[12] What about the "system" of pattern description that simply baptizes frames with proper names (*A* through *F*, in this case) and tells the receiver which frame is up by simply sending '*F*'? This looks much shorter than the bit map until we consider that such a description must be part of an entirely general system. How many proper names will we need to name all possible 900-dot frames? Trivially, the 900-bit binary number, 11111111 . . . . To send the "worst-case" proper name will take exactly as many bits as sending the bit map. This confirms our intuition that proper names are maximally inefficient ways of couching generalizations ("Alf is tall and Bill is tall and . . .").

mal pattern—and hence the "noise" would not "stand out" as noise. There are myriad ways of displaying any 900-bit series of data points, and not many of them would inspire us to concoct an efficient description of the series. Other creatures with different sense organs, or different interests, might readily perceive patterns that were imperceptible to us. The patterns would be *there* all along, but just invisible to *us*.

The idiosyncracy of perceivers' capacities to discern patterns is striking. Visual patterns with axes of vertical symmetry stick out like sore thumbs for us, but if one simply rotates the frame a few degrees, the symmetry is often utterly beyond noticing. And the "perspectives" from which patterns are "perceptible" are not restricted to variations on presentation to the sense modalities. Differences in knowledge yield striking differences in the capacity to pick up patterns. Expert chess players can instantly perceive (and subsequently recall with high accuracy) the total board position in a real game, but are much worse at recall if the same chess pieces are randomly placed on the board, even though to a novice both boards are equally hard to recall.[13] This should not surprise anyone who considers that an expert speaker of English would have much less difficulty perceiving and recalling

> The frightened cat struggled to get loose.

than

> Te ser.ioghehnde t srugfcalde go tgtt ohle

which contains the same pieces, now somewhat disordered. Expert chess players, unlike novices, not only know how to *play* chess; they know how to *read* chess—how to see the patterns at a glance.

A pattern exists in some data—is real—if *there is* a description of the data that is more efficient than the bit map, whether or not anyone can concoct it. Compression algorithms, as general-purpose pattern describers, are efficient ways of transmitting exact copies of frames, such as *A–F*, from one place to another, but our interests often favor a somewhat different goal: transmitting *inexact* copies that nevertheless preserve "the" pattern that is important to us. For some purposes, we need not list the exceptions to bar code, but only transmit the information that the pattern is bar code with *n*% noise. Following this strategy, frames *A* and *C*, though discernibly different under careful inspection, count as *the same pattern*, since what matters to us is that the pattern is bar code with 25% noise, and we do not care which particular noise occurs, only that it occurs.

---

[13] A. D. de Groot, *Thought and Choice in Chess* (The Hague: Mouton, 1965).

Sometimes we are interested in not just ignoring the noise, but eliminating it, improving the pattern in transmission. Copy-editing is a good example. Consider the likely effoct thes santince wull hive hod on tha cupy adutor whu preparis thas monescrupt fur prunteng. *My* interest in this particular instance is that the "noise" be transmitted, not removed, though I actually do not care exactly *which* noise is there.

Here then are three different attitudes we take at various times toward patterns. Sometimes we care about exact description or reproduction of detail, at whatever cost. From this perspective, a real pattern in frame *A* is *bar code with the following exceptions: 7, 8, 11, . . .* .. At other times we care about the noise, but not where in particular it occurs. From this perspective, a real pattern in frame *A* is *bar code with 25% noise*. And sometimes, we simply tolerate or ignore the noise. From this perspective, a real pattern in frame *A* is simply: *bar code*. But is bar code really there in frame A? I am tempted to respond: Look! You can see it with your own eyes. But there is something more constructive to say as well.

When two individuals confront the same data, they may perceive different patterns in them, but since we can have varied interests and perspectives, these differences do not all count as disagreements. Or in any event they should not. If Jones sees pattern α (with n% noise) and Brown sees pattern β (with m% noise) there may be no ground for determining that one of them is right and the other wrong. Suppose they are both using their patterns to bet on the next datum in the series. Jones bets according to the "pure" pattern α, but budgets for n% errors when he looks for odds. Brown does likewise, using pattern β. If both patterns are real, they will both get rich. That is to say, so long as they use their expectation of deviations from the "ideal" to temper their odds policy, they will do better than chance —perhaps very much better.

Now suppose they compare notes. Suppose that α is a simple, easy-to-calculate pattern, but with a high noise rate—for instance, suppose α is bar code as it appears in frame *E*. And suppose that Brown has found some periodicity or progression in the "random" noise that Jones just tolerates, so that β is a much more complicated description of pattern-superimposed-on-pattern. This permits Brown to do better than chance, we may suppose, at predicting when the "noise" will come. As a result, Brown budgets for a lower error rate—say only 5%. "What you call noise, Jones, is actually pattern," Brown might say. "Of course there is still *some* noise in my pattern, but my pattern is better—more real—than yours! Yours is actually just a mere appearance." Jones might well reply that it is all a matter of taste; he notes how hard Brown has to work to calculate predictions, and points to the fact that he is getting just as rich (or maybe richer) by using a simpler, sloppier system and making more bets at good odds than Brown can muster. "My pattern is perfectly real—look how rich I'm getting. If it were an illusion, I'd be broke."

This crass way of putting things—in terms of betting and getting rich—is simply a vivid way of drawing attention to a real, and far from crass, trade-off that is ubiquitous in nature, and hence in folk psychology. Would we prefer an extremely compact pattern description with a high noise ratio or a less compact pattern description with a lower noise ratio? Our decision may depend on how swiftly and reliably we can discern the simple pattern, how dangerous errors are, how much of our resources we can afford to allocate to detection and calculation. These "design decisions" are typically not left to us to make by individual and deliberate choices; they are incorporated into the design of our sense organs by genetic evolution, and into our culture by cultural evolution. The product of this design evolution process is what Wilfrid Sellars[14] calls our *manifest image,* and it is composed of folk physics, folk psychology, and the other pattern-making perspectives we have on the buzzing blooming confusion that bombards us with data. The ontology generated by the manifest image has thus a deeply pragmatic source.[15]

Do these same pragmatic considerations apply to the scientific image, widely regarded as the final arbiter of ontology? Science is supposed to carve nature at the joints—at its *real* joints, of course. Is it permissible in science to adopt a carving system so simple that it makes sense to tolerate occasional misdivisions and consequent mispredictions? It happens all the time. The ubiquitous practice of using idealized models is exactly a matter of trading off reliability and accuracy of prediction against computational tractability. A particularly elegant and handy oversimplification may under some circumstances be irresistible. The use of Newtonian rather than Einsteinian mechanics in most mundane scientific and engineering calculations is an obvious example. A tractable oversimplification may be attractive even in the face of a high error rate; considering inherited traits to be carried by single genes "for" those traits is an example; considering agents in the marketplace to be perfectly rational self-aggrandizers with perfect information is another.

---

[14] *Science, Perception and Reality* (Boston: Routledge & Kegan Paul, 1963).

[15] In "Randomness and Perceived Randomness in Evolutionary Biology," *Synthese,* XLIII (1980): 287–329, William Wimsatt offers a nice example (296): while the insectivorous bird tracks individual insects, the anteater just averages over the ant-infested area; one might say that, while the bird's manifest image quantifies over insects, 'ant' is a mass term for anteaters. See the discussion of this and related examples in my *Elbow Room* (Cambridge: MIT, 1984), pp. 108–110.

III. PATTERNS IN LIFE

The time has come to export these observations about patterns and reality to the controversial arena of belief attribution. The largish leap we must make is nicely expedited by pausing at a stepping-stone example midway between the world of the dot frames and the world of folk psychology: John Horton Conway's Game of Life. In my opinion, every philosophy student should be held responsible for an intimate acquaintance with the Game of Life. It should be considered an essential tool in every thought-experimenter's kit, a prodigiously versatile generator of philosophically important examples and thought experiments of admirable clarity and vividness. In *The Intentional Stance,* I briefly exploited it to make a point about the costs and benefits of risky prediction from the intentional stance,[16] but I have since learned that I presumed too much familiarity with the underlying ideas. Here, then, is a somewhat expanded basic introduction to Life.[17]

Life is played on a two-dimensional grid, such as a checkerboard or a computer screen; it is not a game one plays to win; if it is a game at all, it is solitaire. The grid divides space into square cells, and each cell is either ON or OFF at each moment. Each cell has eight neighbors: the four adjacent cells north, south, east, and west, and the four diagonals: northeast, southeast, southwest, and northwest. Time in the Life world is also discrete, not continuous; it advances in ticks, and the state of the world changes between each tick according to the following rule:

> Each cell, in order to determine what to do in the next instant, counts how many of its eight neighbors is ON at the present instant. If the answer is exactly two, the cell stays in its present state (ON or OFF) in the next instant. If the answer is exactly three, the cell is ON in the next instant whatever its current state. Under all other conditions the cell is OFF.

The entire physics of the Life world is captured in that single, unexceptioned law. [While this is the fundamental law of the "physics" of the Life world, it helps at first to conceive this curious physics in biological terms: think of cells going ON as births, cells going OFF as deaths, and succeeding instants as generations. Either overcrowding (more than three inhabited neighbors) or isolation (less than two inhabited neighbors) leads to death.] By the scrupulous application of this single law, one can predict with perfect accuracy the next instant of any configuration of ON and OFF cells, and the instant after that, and so forth. In other words, the Life world is a toy world that perfectly instantiates Laplace's vision of determinism: given the state description of this world at an instant, we finite observers can perfectly predict the future instants by the simple application of our one law of physics. Or, in my terms, when we adopt the physical stance toward a configuration in the Life world, our powers of prediction are perfect: there is no noise, no uncertainty, no probability less than one. Moreover, it follows from the two-dimensionality of the Life world that nothing is hidden from view. There is no backstage; there are no hidden variables; the unfolding of the physics of objects in the Life world is directly and completely visible.

---

[16] *The Intentional Stance,* pp. 37–9.

[17] Martin Gardner introduced the Game of Life to a wide audience in two columns in *Scientific American* in October, 1970, and February, 1971. William Poundstone, *The Recursive Universe: Cosmic Complexity and the Limits of Scientific Knowledge* (New York: Morrow, 1985), is an excellent exploration of the game and its philosophical implications. Two figures from Poundstone's book are reproduced, with kind permission from the author and publisher, on pp. 39 & 40.

There are computer simulations of the Life world in which one can set up configurations on the screen and then watch them evolve according to the single rule. In the best simulations, one can change the scale of both time and space, alternating between close-up and bird's-eye view. A nice touch added to some color versions is that ON cells (often just called pixels) are color-coded by their age; they are born blue, let us say, and then change color each generation, moving through green to yellow to orange to red to brown to black and then staying black unless they die. This permits one to see at a glance how old certain patterns are, which cells are co-generational, where the birth action is, and so forth.[18]

One soon discovers that some simple configurations are more interesting than others. In addition to those configurations which never change—the "still lifes" such as four pixels in a square—and those which evaporate entirely—such as any long diagonal line segment, whose two tail pixels die of isolation each instant until the line disappears entirely—there are configurations with all manner of periodicity. Three pixels in a line make a simple flasher, which becomes three pixels in a column in the next instant, and reverts to three in a line in the next, ad infinitum, unless some other configuration encroaches. Encroachment is what makes Life interesting:

---

[18] Poundstone, *op. cit.*, provides simple BASIC and IBM-PC assembly language simulations you can copy for your own home computer, and describes some of the interesting variations.

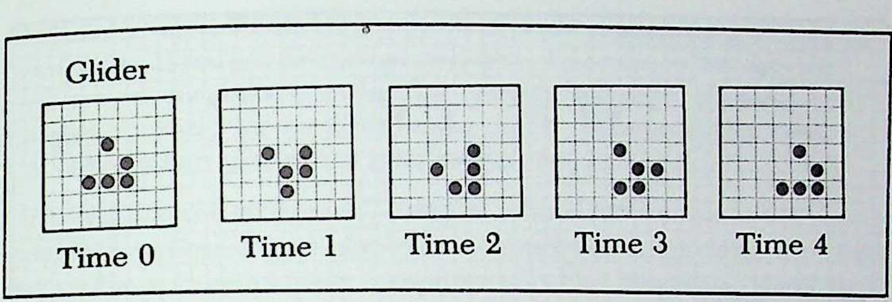

Figure 2: (from Poundstone, *op. cit.*).

among the periodic configurations are some that swim, amoeba-like, across the plane. The simplest is the *glider,* the five-pixel configuration shown taking a single stroke to the southeast in figure 2. Then there are the eaters, the puffer trains, and space rakes, and a host of other aptly named denizens of the Life world that emerge in the ontology of a new level, analogous to what I have called the design level. This level has its own language, a transparent foreshortening of the tedious descriptions one could give at the physical level. For instance:

> An eater can eat a glider in four generations. Whatever is being consumed, the basic process is the same. A bridge forms between the eater and its prey. In the next generation, the bridge region dies from overpopulation, taking a bite out of both eater and prey. The eater then repairs itself. The prey usually cannot. If the remainder of the prey dies out as with the glider, the prey is consumed (*ibid.*, p. 38).

Note that there has been a distinct ontological shift as we move between levels; whereas at the physical level there is no motion, and the only individuals, cells, are defined by their fixed spatial location, at this design level we have the motion of persisting objects; it is one and the same glider that has moved southeast in figure 2, changing shape as it moves, and there is one less glider in the world after the eater has eaten it in figure 3. (Here is a warming-up exercise for what is to follow: should we say that there is *real* motion in the Life world, or only *apparent* motion? The flashing pixels on the computer screen are a paradigm case, after all, of what a psychologist would call apparent motion. Are there *really* gliders that move, or are there just patterns of cell state that move? And if we opt for the latter, should we say at least that these moving patterns are real?)

Notice, too, that at this level one proposes generalizations that require 'usually' or 'provided nothing encroaches' clauses. Stray bits of debris from earlier events can "break" or "kill" one of the objects

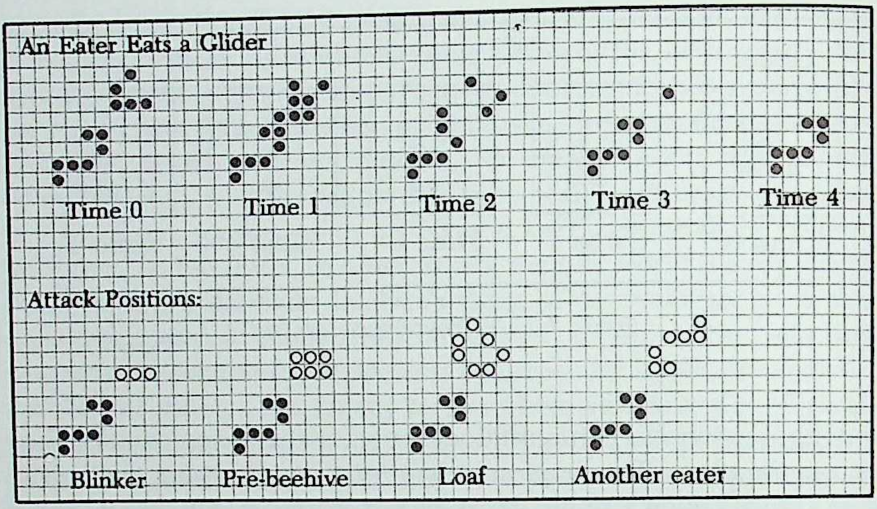

Figure 3: (from Poundstone, *op. cit.*).

in the ontology at this level; their *salience as real things* is considerable, but not guaranteed. To say that their salience is considerable is to say that one can, with some small risk, ascend to this design level, adopt its ontology, and proceed to predict—sketchily and riskily—the behavior of larger configurations or systems of configurations, without bothering to compute the physical level. For instance, one can set oneself the task of designing some interesting supersystem out of the "parts" that the design level makes available. Surely the most impressive triumph of this design activity in the Life world is the proof that a working model of a universal Turing machine can in principle be constructed in the Life plane! Von Neumann had already shown that in principle a two-dimensional universal Turing machine could be constructed out of cellular automata, so it was "just" a matter of "engineering" to show how, in principle, it could be constructed out of the simpler cellular automata defined in the Life world. Glider streams can provide the tape, for instance, and the tape reader can be some huge assembly of eaters, gliders, and other bits and pieces. What does this huge Turing machine look like? Poundstone calculates that the whole construction, a self-reproducing machine incorporating a universal Turing machine, would be on the order of 10^{13} pixels.

> Displaying a 10^{13}-pixel pattern would require a video screen about 3 million pixels across at least. Assume the pixels are 1 millimeter square (which is very high resolution by the standards of home computers). Then the screen would have to be 3 kilometers (about two miles) across. It would have an area about six times that of Monaco.

> Perspective would shrink the pixels of a self-reproducing pattern to invisibility. If you got far enough away from the screen so that the entire pattern was comfortably in view, the pixels (and even the gliders, eaters and guns) would be too tiny to make out. A self-reproducing pattern would be a hazy glow, like a galaxy (*ibid.*, pp. 227–8).

Now, since the universal Turing machine can compute any computable function, it can play chess—simply by mimicking the program of any chess-playing computer you like. Suppose, then, that such an entity occupies the Life plane, playing chess against itself. Looking at the configuration of dots that accomplishes this marvel would almost certainly be unilluminating to anyone who had no clue that a configuration with such powers could exist. But from the perspective of one who had the hypothesis that this huge array of black dots was a chess-playing computer, enormously efficient ways of predicting the future of that configuration are made available. As a first step one can shift from an ontology of gliders and eaters to an ontology of symbols and machine states, and, adopting this higher design stance toward the configuration, predict its future *as* a Turing machine. As a second and still more efficient step, one can shift to an ontology of chess-board positions, possible chess moves, and the grounds for evaluating them; then, adopting the intentional stance toward the configuration, one can predict its future *as* a chess player performing intentional actions—making chess moves and trying to achieve checkmate. Once one has fixed on an interpretation scheme, permitting one to say which configurations of pixels count as which symbols (either, at the Turing machine level, the symbols '0' or '1', say, or at the intentional level, '*QxBch*' and the other symbols for chess moves), one can use the interpretation scheme to predict, for instance, that the next configuration to emerge from the galaxy will be such-and-such a glider stream (the symbols for '*RxQ*', say). There is risk involved in either case, because the chess program being run on the Turing machine may be far from perfectly rational, and, at a different level, debris may wander onto the scene and "break" the Turing machine configuration before it finishes the game.

In other words, real but (potentially) noisy patterns abound in such a configuration of the Life world, there for the picking up if only we are lucky or clever enough to hit on the right perspective. They are not *visual* patterns but, one might say, *intellectual* patterns. Squinting or twisting the page is not apt to help, while posing fanciful interpretations (or what W. V. Quine would call "analytical hypotheses") may uncover a goldmine. The opportunity confronting the observer of such a Life world is analogous to the opportunity confronting the cryptographer staring at a new patch of cipher text, or the opportunity confronting the Martian, peering through a telescope at the Superbowl Game. If the Martian hits on the intentional stance—or folk psychology—as the right level to look for pattern, shapes will readily emerge through the noise.

IV. THE REALITY OF INTENTIONAL PATTERNS

The scale of compression when one adopts the intentional stance toward the two-dimensional chess-playing computer galaxy is stupendous: it is the difference between figuring out in your head what white's most likely (best) move is versus calculating the state of a few trillion pixels through a few hundred thousand generations. But the scale of the savings is really no greater in the Life world than in our own. Predicting that someone will duck if you throw a brick at him is easy from the folk-psychological stance; it is and will always be intractable if you have to trace the photons from brick to eyeball, the neurotransmitters from optic nerve to motor nerve, and so forth.

For such vast computational leverage one might be prepared to pay quite a steep price in errors, but in fact one belief that is shared by all of the representatives on the spectrum I am discussing is that "folk psychology" provides a description system that permits highly reliable prediction of human (and much nonhuman) behavior.[19] They differ in the explanations they offer of this predictive prowess, and the implications they see in it about "realism."

For Fodor, an industrial-strength Realist, beliefs and their kin would not be real unless the pattern dimly discernible from the perspective of folk psychology could also be discerned (more clearly, with less noise) as a pattern of structures in the brain. The pattern would have to be discernible from the different perspective provided by a properly tuned *syntactoscope* aimed at the purely formal (nonsemantic) features of Mentalese terms written in the brain. For Fodor, the pattern seen through the noise by everyday folk psychologists would tell us nothing about reality, unless it, and the noise, had the following sort of explanation: what we discern from the perspective of folk psychology is the net effect of two processes: an ulterior, hidden process wherein the pattern exists quite pure, overlaid, and partially obscured by various intervening sources of noise: performance errors, observation errors, and other more or less random obstructions. He might add that the interior belief-producing process was in this respect *just* like the process responsible for the creation of frames *A–F*. If you were permitted to peer behind the scenes at the program I devised to create the frames, you would see, clear as a bell, the perfect bar-code periodicity, with the noise thrown on afterward like so much salt and pepper.

---

[19] To see that the opposite poles share this view, see Fodor, *Psychosemantics* (Cambridge: MIT, 1987), ch. 1, "Introduction: the Persistence of the Attitudes"; and Paul Churchland, *Scientific Realism and the Plasticity of Mind* (New York: Cambridge, 1979), esp. p. 100: "For the P-theory [folk psychology] is in fact a marvelous intellectual achievement. It gives its possessor an explicit and systematic insight into the behaviour, verbal and otherwise, of some of the most complex agents in the environment, and its overall prowess in that respect remains unsurpassed by anything else our contemporary theoretical efforts have produced."

This is often the explanation for the look of a data set in science, and Fodor may think that it is either the only explanation that can ever be given, or at any rate the only one that makes any sense of the success of folk psychology. But the rest of us disagree. As G. E. M. Anscombe[20] put it in her pioneering exploration of intentional explanation, "if Aristotle's account [of reasoning using the practical syllogism] were supposed to describe actual mental processes, it would in general be quite absurd. The interest of the account is that it describes an order which is there whenever actions are done with intentions . . ." (*ibid.*, p. 80).

But how *could* the order be there, so visible amidst the noise, if it were not the direct outline of a concrete orderly process in the background? Well, it *could* be there thanks to the statistical effect of very many concrete minutiae producing, as if by a hidden hand, an approximation of the "ideal" order. Philosophers have tended to ignore a variety of regularity intermediate between the regularities of planets and other objects "obeying" the laws of physics and the regularities of rule-following (that is, rule-*consulting*) systems.[21] These intermediate regularities are those which are preserved under selection pressure: the regularities dictated by principles of good design and hence homed in on by self-designing systems. That is, a "rule of thought" may be much more than a mere regularity; it may be a *wise* rule, a rule one would design a system by if one were a system designer, and hence a rule one would expect self-designing systems to "discover" in the course of settling into their patterns of activity. Such rules no more need be explicitly represented than do the principles of aerodynamics that are honored in the design of birds' wings.[22]

---

[20] *Intention* (New York: Blackwell, 1957).

[21] A notable early exception is Sellars, who discussed the importance of just this sort of regularity in "Some Reflections on Language Games," *Philosophy of Science*, XXI (1954): 204–228. See especially the subsection of this classic paper, entitled "Pattern Governed and Rule Obeying Behavior," reprinted in Sellars's *Science, Perception and Reality*, pp. 324–7.

[22] Several interpreters of a draft of this article have supposed that the conclusion I am urging here is that beliefs (or their contents) are *epiphenomena* having no causal powers, but this is a misinterpretation traceable to a simplistic notion of causation. If one finds a predictive pattern of the sort just described one has *ipso facto* discovered a causal power, a difference in the world that makes a subsequent

The contrast between these different sorts of pattern-generation processes can be illustrated. The frames in figure 1 were created by a hard-edged process (ten black, ten white, ten black, . . .) obscured by noise, while the frames in figure 4 were created by a process almost the reverse of that: the top frame shows a pattern created by a normal distribution of black dots around means at $x = 10, 30, 50, 70$, and 90 (rather like Mach bands or interference fringes); the middle and bottom frames were created by successive applications of a very simple contrast enhancer applied to the top frame: a vertical slit "window" three pixels high is thrown randomly onto the frame; the pixels in the window vote, and majority rules. This gradually removes the salt from the pepper and the pepper from the salt, creating "artifact" edges such as those discernible in the bottom frame. The effect would be more striking at a finer pixel scale, where the black merges imperceptibly through grays to white but I chose to keep the scale at the ten-pixel period of bar code. I do not mean to suggest that it is impossible to tell the patterns in figure 4 from the patterns in figure 1. Of course it is possible; for one thing, the process that produced the frames in figure 1 will almost always show edges at exactly 10, 20, 30, . . . and almost never at 9, 11, 19, 21, . . . while there is a higher probability of these "displaced" edges being created by the process of figure 4 (as a close inspection of figure 4 reveals). Fine tuning could of course reduce these probabilities, but that is not my point. My point is that *even if* the evidence is substantial that the discernible pattern is produced by one process rather than another, it can be rational to ignore those differences and use the simplest pattern description (e.g., *bar code*) as one's way of organizing the data.

Fodor and others have claimed that an interior language of thought is the best explanation of the hard edges visible in "propositional attitude psychology." Churchland and I have offered an alternative explanation of these edges, an explanation for which the

---

difference testable by standard empirical methods of variable manipulation. Consider the crowd-drawing power of a sign reading "Free Lunch" placed in the window of a restaurant, and compare its power in a restaurant in New York to its power in a restaurant in Tokyo. The intentional level is obviously the right level at which to predict and explain such causal powers; the sign more reliably produces a particular belief in one population of perceivers than in the other, and variations in the color of typography of the sign are not as predictive of variations in crowd-drawing power as are variations in (perceivable) meaning. The fact that the regularities on which these successful predictions are based are efficiently capturable (only) in intentional terms and are not derived from "covering laws" does not show that the regularities are not "causal"; it just shows that philosophers have often relied on pinched notions of causality derived from exclusive attention to a few examples drawn from physics and chemistry. Smith has pointed out to me that here I am echoing Aristotle's claim that his predecessors had ignored final causes.

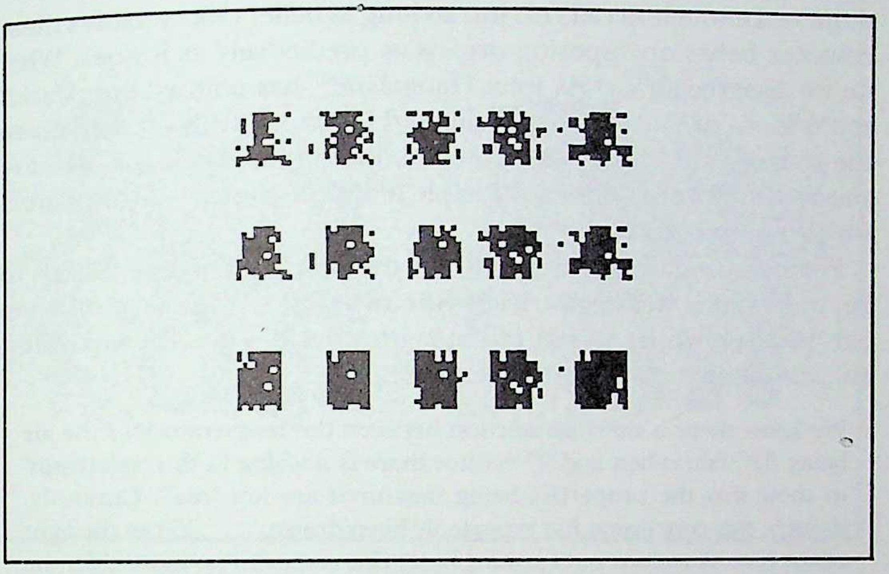

Figure 4

process that produced the frames in figure 4 is a fine visual metaphor. The process that produces the data of folk psychology, we claim, is one in which the multidimensional complexities of the underlying processes are projected *through linguistic behavior,* which creates an appearance of definiteness and precision, thanks to the discreteness of words.[23] As Churchland[24] puts it, a person's declarative utterance is a "one-dimensional *projection*—through the compound lens of Wernicke's and Broca's areas onto the idiosyncratic surface of the speaker's language—a one-dimensional projection of a four- or five-dimensional 'solid' that is an element in his true kinematic state" (*ibid.,* p. 85).

Fodor's industrial-strength Realism takes beliefs to be things in the head—just like cells and blood vessels and viruses. Davidson and I both like Churchland's alternative idea of propositional-attitude statements as indirect "measurements" of a reality diffused in the behavioral dispositions of the brain (and body).[25] We think beliefs are quite real enough to call real just so long as belief talk measures these complex behavior-disposing organs as predictively as it does. What do we disagree about? As John Haugeland[26] has pointed out, Davidson is more of a realist than I am, and I have recently tracked down the source of this disagreement to a difference of opinion we have about the status of Quine's principle of indeterminacy of translation, which we both accept.

---

[23] See my discussion of the distinction between beliefs and (linguistically infected) *opinions;* in *Brainstorms* (Montgomery, VT: Bradford, 1978), ch. 16, and in "The Illusions of Realism," in *The Intentional Stance,* pp. 110–6.

[24] "Eliminative Materialism and the Propositional Attitudes," this JOURNAL, LXXVIII, 2 (February 1981): 67–90, esp. p. 85.

[25] Churchland introduces the idea in *Scientific Realism and the Plasticity of Mind,* pp. 100–7. My adoption of the idea was in "Beyond Belief," in A. Woodfield, ed., *Thought and Object* (New York: Oxford, 1982), repr. as ch. 5 of *The Intentional Stance.* Davidson's guarded approval is expressed in "What is Present to the Mind?" read at the Sociedad Filosófica Ibero Americana meeting in Buenos Aires, 1989.

For Davidson, the principle is not the shocker it is often taken to be; in fact, it is well-nigh trivial—the two different translation manuals between which no fact of the matter decides are like two different scales for measuring temperature.

> We know there is no contradiction between the temperature of the air being 32° fahrenheit and 0° celsius; there is nothing in this 'relativism' to show that the properties being measured are not 'real'. Curiously, though, this conclusion has repeatedly been drawn. . . . Yet in the light of the considerations put forward here, this comes to no more than the recognition that more than one set of one person's utterances might be equally successful in capturing the contents of someone else's thoughts or speech. Just as numbers can capture all the empirically significant relations among weights or temperatures in infinitely many ways, so one person's utterances can capture all the significant features of another person's thoughts and speech in different ways. This fact does not challenge the 'reality' of the attitudes or meanings thus reported.[27]

On Davidson's view, no substantive disagreements emerge from a comparison of the two description schemes, and so they can quite properly be viewed as competing descriptions of the same reality.

I think this is a flawed analogy. A better one is provided by the example of "rival" descriptions of patterns-with-noise. Consider two rival intentional interpretations of a single individual; they agree on the general shape of this individual's collection of beliefs (and desires, etc), but because of their different idealizations of the pattern, they do not agree point-for-point. Recalling a famous analogy of Quine's[28] and extending it beyond radical translation to radical interpretation (as Davidson and I both wish to do), we get the image in Figure 5.

To the left we see Brown's intentional interpretation of Ella; to the right, Jones's interpretation. Since these are intentional interpreta-

---

[26] See the discussion of Haugeland's views in the last chapter of *The Intentional Stance*, "Mid-Term Examination: Compare and Contrast," pp. 348–9.

[27] Davidson, "What is Present to the Mind?" (ms.), p. 10.

[28] "Different persons growing up in the same language are like different bushes trimmed and trained to take the shape of identical elephants. The anatomical details of twigs and branches will fulfill the elephantine form differently from bush to bush, but the overall outward results are the same." *Word and Object* (Cambridge: MIT, 1960), p. 8.

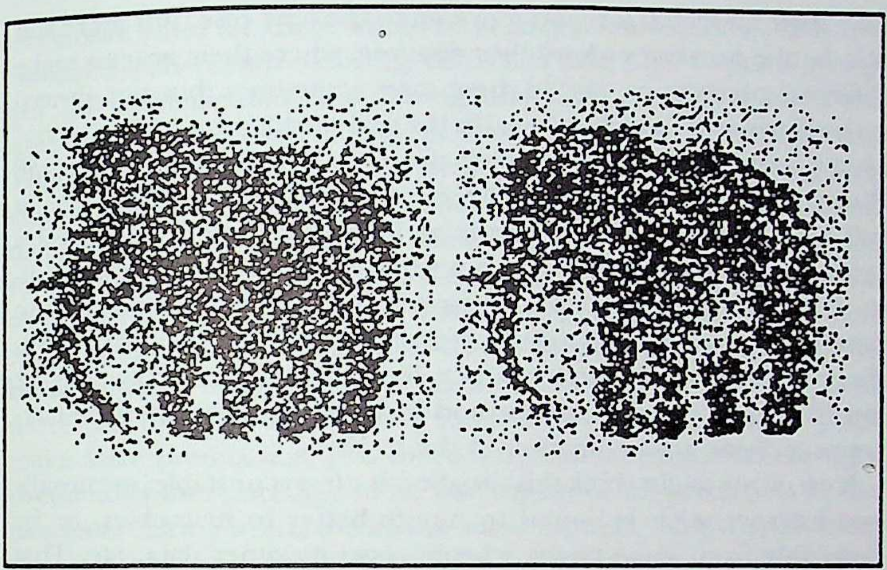

Figure 5

tions, the pixels or data points represent beliefs and so forth, not (for instance) bits of bodily motion or organs or cells or atoms, and since these are rival intentional interpretations of a single individual, the patterns discerned are not statistical averages (e.g., "Democrats tend to favor welfare programs") but personal cognitive idiosyncracies (e.g., "She thinks she should get her queen out early"). Some of the patterns may indeed be simple observed periodicities (e.g., "Ella wants to talk about football on Mondays") but we are to understand the pattern to be what Anscombe called the "order which is there" in the rational coherence of a person's set of beliefs, desires, and intentions.

Notice that here the disagreements can be substantial—at least before the fact: when Brown and Jones make a series of predictive bets, they will not always make the same bet. They may *often* disagree on what, according to their chosen pattern, will happen next. To take a dramatic case, Brown may predict that Ella will decide to kill herself; Jones may disagree. This is not a trivial disagreement of prediction, and in principle this momentous difference may emerge in spite of the overall consonance of the two interpretations.

Suppose, then, that Brown and Jones make a series of predictions of Ella's behavior, based on their rival interpretations. Consider the different categories that compose their track records. First, there are the occasions where they agree and are right. Both systems look good from the vantage point of these successes. Second, there are the occasions where they agree and are wrong. Both chalk it up to noise, take their budgeted loss and move on to the next case. But there will also be the occasions where they disagree, where their systems make different predictions, and in these cases sometimes (but not always) one will win and the other lose. (In the real world, predictions are not always from among binary alternatives, so in many cases they will disagree and both be wrong.) When one wins and the other loses, it will look to the myopic observer as if one "theory" has scored a serious point against the other, but when one recognizes the possibility that both may chalk up such victories, and that there may be no pattern in the victories which permits either one to improve his theory by making adjustments, one sees that local triumphs may be insufficient to provide any ground in reality for declaring one account a closer approximation of the truth.

Now, some might think this situation is *always* unstable; eventually one interpretation is bound to ramify better to new cases, or be deducible from some larger scheme covering other data, etc. That might be true in many cases, but—and this, I think, is the central point of Quine's indeterminacy thesis—it need not be true in all. *If* the strategy of intentional-stance description is, as Quine says, a "dramatic idiom" in which there is ineliminable use of idealization, and if Fodor's industrial-strength Realism is thus not the correct explanation of the reliable "visibility" of the pattern, such radical indeterminacy is a genuine and stable possibility.

This indeterminacy will be most striking in such cases as the imagined disagreement over Ella's suicidal mindset. If Ella does kill herself, is Brown shown to have clearly had the better intentional interpretation? Not necessarily. When Jones chalks up his scheme's failure in this instance to a bit of noise, this is no more ad hoc or unprincipled than the occasions when Brown was wrong about whether Ella would order the steak not the lobster, and chalked those misses up to noise. This is not at all to say that an interpretation can never be shown to be just wrong; there is plenty of leverage within the principles of intentional interpretation to refute particular hypotheses—for instance, by forcing their defense down the path of Pickwickian explosion ("You see, she didn't believe the gun was loaded because she thought that those bullet-shaped things were chocolates wrapped in foil, which was just a fantasy that occurred to her because . . . ."). It *is* to say that there could be two interpretation schemes that were reliable and compact predictors over the long run, but that nevertheless disagreed on crucial cases.

It might seem that in a case as momentous as Ella's intention to kill herself, a closer examination of the details just prior to the fatal moment (if not at an earlier stage) would have to provide additional support for Brown's interpretation at the expense of Jones's interpretation. After all, there would be at least a few seconds—or a few hundred milliseconds—during which Ella's decision to pull the trigger got implemented, and during that brief period, at least, the evidence would swing sharply in favor of Brown's interpretation. That is no doubt true, and it is *perhaps* true that had one gone into enough detail earlier, all this last-second detail could have been predicted—but to have gone into *those* details earlier would have been to drop down from the intentional stance to the design or physical stances. From the intentional stance, these determining considerations would have been invisible to both Brown and Jones, who were both prepared to smear over such details as noise in the interests of more practical prediction. Both interpreters concede that they will make false predictions, and moreover, that when they make false predictions there are apt to be harbingers of misprediction in the moments during which the *dénouement* unfolds. Such a brief swing does not constitute refutation of the interpretation, any more than the upcoming misprediction of behavior does.

How, then, does this make me less of a realist than Davidson? I see that there could be two different systems of belief attribution to an individual which differed *substantially* in what they attributed—even in yielding substantially different predictions of the individual's future behavior—and yet where no deeper fact of the matter could establish that one was a description of the individual's *real* beliefs and the other not. In other words, there could be two different, but equally real, patterns discernible in the noisy world. The rival theorists would not even agree on which parts of the world were pattern and which were noise, and yet nothing deeper would settle the issue.[29] The choice of a pattern would indeed be up to the observer, a matter to be decided on idiosyncratic pragmatic grounds. I myself do not see any feature of Davidson's position that would be a serious obstacle to his shifting analogies and agreeing with me. But then he would want to grant that indeterminacy is not such a trivial matter after all.[30]

What then is Rorty's view on these issues? Rorty wants to deny that any brand of "realism" could *explain* the (apparent?) success of the intentional stance. But since we have already joined Fine and set aside the "metaphysical" problem of realism, Rorty's reminding us of this only postpones the issue. Even someone who has transcended the scheme/content distinction and has seen the futility of correspondence theories of truth must accept the fact that *within* the natural ontological attitude we sometimes explain success by correspondence: one does better navigating off the coast of Maine when one uses an up-to-date nautical chart than one does when one uses a road map of Kansas. Why? Because the former accurately represents the hazards, markers, depths, and coastlines of the Maine coast, and the latter does not. Now why does one do better navigating the shoals of interpersonal relations using folk psychology than using astrology? Rorty might hold that the predictive "success" we folk-psychology players relish is itself an artifact, a mutual agreement engendered by the egging-on or consensual support we who play this game provide each other. He would grant that the game has no rivals in popularity, due—in the opinion of the players—to the power it gives them to understand and anticipate the animate world. But he would refuse to endorse this opinion. How, then, would he distinguish this popularity from the popularity among a smaller coterie of astrology?[31] It is undeniable that astrology provides its adherents with a highly articulated system of patterns that they *think* they see in the events of the world. The difference, however, is that no one has ever been able to get rich by betting on the patterns, but only by selling the patterns to others.

---

[29] Cf. "The Abilities of Men and Machines," in *Brainstorms*, where I discuss two people who agree exactly on the future behavior of some artifact, but impose different Turing-machine interpretations of it. On both interpretations, the machine occasionally "makes errors" but the two interpreters disagree about which cases are the errors. (They disagree about which features of the object's behavior count as signal and which as noise.) Which Turing machine is it really? This question has no answer.

[30] Andrej Zabludowski seems to me to have overlooked this version of indeterminacy in "On Quine's Indeterminacy Doctrine," *Philosophical Review*, XCVIII (1989): 35–64.

Rorty would have to claim that this is not a significant difference; the rest of us, however, find abundant evidence that our allegiance to folk psychology as a predictive tool can be defended in coldly objective terms. We agree that there is a real pattern being described by the terms of folk psychology. What divides the rest of us is the nature of the pattern, and the ontological implications of that nature.

Let us finally consider Churchland's eliminative materialism from this vantage point. As already pointed out, he is second to none in his appreciation of the power, to date, of the intentional stance as a strategy of prediction. Why does he think that it is nevertheless doomed to the trash heap? Because he anticipates that neuroscience will eventually—perhaps even soon—discover a pattern that is so clearly superior to the noisy pattern of folk psychology that everyone will readily abandon the former for the latter (except, perhaps, in the rough-and-tumble of daily life). This might happen, I suppose. But

---

[31] Cf. my comparison of "the astrological stance" to the intentional stance, *The Intentional Stance*, p. 16.

Churchland here is only playing a hunch, a hunch that should not be seen to gain plausibility from reflections on the irresistible forward march of science. For it is not enough for Churchland to suppose that in principle, neuroscientific levels of description will explain more of the variance, predict more of the "noise" that bedevils higher levels. This is, of course, bound to be true in the limit—if we descend all the way to the neurophysiological "bit map." But as we have seen, the trade-off between ease of use and immunity from error for such a cumbersome system may make it profoundly unattractive.[32] If the "pattern" is scarcely an improvement over the bit map, talk of eliminative materialism will fall on deaf ears—just as it does when radical eliminativists urge us to abandon our ontological commitments to tables and chairs. A truly general-purpose, robust system of pattern description more valuable than the intentional stance is not an impossibility, but anyone who wants to bet on it might care to talk to me about the odds they will take.

What does all this show? Not that Fodor's industrial-strength Realism must be false, and not that Churchland's eliminative materialism must be false, but just that both views are gratuitously strong forms of materialism—presumptive theses way out in front of the empirical support they require. Rorty's view errs in the opposite direction, ignoring the impressive empirical track record that distinguishes the intentional stance from the astrological stance. Davidson's intermediate position, like mine, ties reality to the brute existence of pattern, but Davidson has overlooked the possibility of two or more *conflicting* patterns being superimposed on the same data—a more radical indeterminacy of translation than he had supposed possible. Now, once again, is the view I am defending here a sort of instrumentalism or a sort of realism? I think that the view itself is clearer than either of the labels, so I shall leave that question to anyone who stills find illumination in them.

DANIEL C. DENNETT

Tufts University

---

[32] As I have put it, physical-stance predictions trump design-stance predictions, which trump intentional-stance predictions—but one pays for the power with a loss of portability and a (usually unbearable) computational cost.

BOOK REVIEWS

The Nature of Mental Things. ARTHUR W. COLLINS. Notre Dame: Notre Dame University Press, 1987. xviii, 178 p. Cloth $22.95.

Is philosophy prone to fashion? And if so, can we think of a philosophical position like a dress that we can store in the attic to await the time when the current fashion makes it acceptable to wear again? Reading *The Nature of Mental Things* puts one in mind of a good rummage around in the attic. The things one finds are not so very old—just a generation or so. Some of what one finds is a little dusty, and some misses the point of what has come since, but much of what one finds makes one reflect upon the life that went before; if one is sufficiently open, it can teach one much about the present.

In his recent book, Arthur W. Collins is concerned with the nature of mind. He writes globally of "the mind," but his real concern is with beliefs, desires, and visual experiences. He aims to show that "there is less 'to the mind' than we philosophers have supposed" (43). He criticizes at length identity theorists who, he suggests, may be either materialists or dualists. The functionalist, who claims that mental states are realized in states of a body or of a soul, is attacked by association. Those philosophers who, in accounting for propositional attitudes, invoke the notion of representation, of languages of thought, of information processing, or of computation are accused of simply turning out variations on a mistaken theme. The mistake is in the assumption that the mind is an inner thing. Heavily influenced by Ludwig Wittgenstein, Collins aims to diagnose how the mistake comes to be made and to dispel the illusion.

Collins insists that common-sense thinking about the mind is adulterated by theory, and he attempts to pick apart the strands. Here are some pretheoretical (and acceptable) reflections: I have beliefs and so do others; some beliefs are true and some are false; the mind is private—that is, most people know about their own mental states much of the time and can tell others about these states with a certain authority; there are states of consciousness—for example, people who see have visual experiences. Here is what philosophical theory (mistakenly) adds to these basic reflections: the mind is an object of privileged access via infallible introspective consciousness; beliefs are neural states or functional states or states of an immortal soul; there is a phenomenon of consciousness to which we attempt to refer by

0022-362X/91/8801/52-6

such expressions as 'immediate experience' or 'inner goings on'. What theory adds to common sense, and where our thinking about the mind goes awry, is in the idea that "beliefs, desires, and other mental things are inner states or processes" (2). Despite leaning heavily on common sense, Collins claims not to rest his case on intuition but on the fact that inner-state theories cannot do the job required of them. It is unclear whether Collins thinks we need a better theory or whether he thinks that the real problem is that we attempt to theorize at all.

Philosophical thinking about consciousness provides a clear illustration of how we go wrong. It starts with the idea of the phenomena of consciousness. It is already on the wrong footing because, claims Collins, phenomena are what a theory must save (64). Its next step is to convert the phenomena into objects of consciousness. Finally, the phenomena, the objects, are identified with either a Cartesian immaterial soul or a material body. (Here we see that Cartesian dualism and materialism are just two sides of the same coin—a point much emphasized by Collins.) The point of the identification is to give an account of the objects of consciousness. But, argues Collins, an account of the objects of consciousness is one thing, an account of consciousness is another. Collins suggests that inner-state theories may be "rescued" by the following move: collapse the object of consciousness into consciousness itself. Thus we are presented with the idea that the objects of consciousness are luminous, they are intrinsically conscious. But now we need some account of this luminosity and, claims Collins, none is forthcoming. Collins concludes that inner-state theories cannot do the job they are required to do. But perhaps we are only entitled to conclude that, as they now stand, inner-state theories do not do the job required.

According to Collins, once we adopt this theory and latch onto the idea of an inner reality, we are doomed to shuffle back and forth between dualism and materialism. We must exorcise the theory. The concept of mind is "outer-oriented" (175). In this respect, analytical behaviorism got things right. But are we not now in danger of shuffling back and forth between the inner and the outer? Having turned our backs on the idea of the mind as an *inner* reality, are we now to swallow the idea (the theory?) of the mind as an *outer* reality? Collins is much less forthcoming here (although his tolerance of analytical behaviorism may be indicative). His attack on eliminativism as a "lunatic philosophy of mind" leads one to conclude that it is not the reality of the mental so much as its innerness that is being rejected.

The suggestion that the mind may not be wholly inner but *also* inner is not considered. One is left wondering whether things would not have looked different if Collins had chosen to concentrate less on the case of visual experience and instead had considered sensations like pain. (A curious footnote appears on page 20 where Collins claims to agree with Wittgenstein that the mental is not an inner state or process yet to disagree with the same over the case of pain. Pains, claims Collins, *are* inner realities.) And even if one does look to visual experience, it is unclear why "his seeing of the wax" is a particularly problematic accompaniment of "he sees the wax." Collins does not accuse these nominalizations of tempting us to repeat after Franz Brentano that inner perception is the only real perception. Direct realism is compatible with the acceptance of introspective awareness. Nor does introspective awareness demand infallibility. One is left to understand that the real problem here is that nominalization tempts one to inner objects, and inner objects tempt one to identifications that leave all that matters unexplained. But, once again, this only shows that the theory has more work to do, not that it is doomed to fail.

Collins discusses Thomas Nagel's work as a case in point: talk of the "subjectivity of experience," of "what it is like," and of "points of view" all succumbs to the temptation of the inner. The fact that Nagel rejects all current forms of reductive materialism but holds out the hope of some possible future reduction only serves to strengthen Collins's point. But there is an aspect of Nagel's work which Collins overlooks: the mental according to Nagel is not just subjective, it is not objective. That is to say, it cannot be understood from an impersonal or detached perspective. (Latterly, Nagel has distinguished physical from mental objectivity—the mind has a *kind* of objectivity, mental objectivity. This is an acknowledgment that there are minds other than our own, but it is also an acknowledgment that, from the outside, we can achieve only an incomplete conception of these minds.) Once we accept that the mind is not objective, cannot be so understood, we should be less inclined simply to move from the mental as inner object to the mental as outer object. It may not be the idea of the innerness of the mental which causes problems so much as that that which is subjective is mistaken for that which is objective. And once we appreciate this, it becomes clear that Cartesian dualism is as much of a mistake as current forms of materialism. The latter is a point much emphasized by Nagel—although Collins writes as if this parallel were overlooked by Nagel (cf. 98, 99).

For a proper rejection of mind-body identifications, Collins turns from Nagel's work to Saul Kripke's. Collins wants to push Kripke's conclusion one step further: once we find that theoretical identifications of visual experience (Collins's example) fail, we ought to reject the very assumption on which they rest: the idea that visual experiences are inner items that call for an identification. We ought to reject the phenomena of visual experience. Perhaps Collins could be clearer about the extent of his divergence from Kripke. The latter after all makes much of a "Cartesian intuition." And there is another difference. Kripke aims to shift a certain burden onto those who claim an identity: explain the intuition of contingency showing it to be an illusion or a mistake. Kripke considers and rejects an explanation that requires that we distinguish mental phenomena and our experiences of them. (The rejection of this explanation does seem even more obvious for Collins's case of visual experience than for Kripke's case of pain.) Collins follows Kripke here, yet Collins concludes that "all theoretical identifications of the form 'visual experiences are . . .' fail" (117). Kripke merely concludes that "the usual moves and analogies are not available to solve the problems of the identity theorist," and he adds that this "is, of course, no proof that no moves are available."[1] It remains open to the identity theorist, then, to come up with some other explanation of the intuition. And some identity theorists believe they have such an alternative explanation—albeit one that relies on making the move from type- to token-identity theories. Token-identity theories are only briefly discussed by Collins—presumably because all identity theories share the same mistaken assumption of inner phenomena.

Collins suggests that the identity theories' real interest is in token mental events and that the move from these to types depends upon the discovery of psychophysical laws. He cites Donald Davidson as an example of someone whose pessimism over such a discovery leaves him embracing only a token-identity theory. Collins's diagnosis, however, does leave us wondering why there is such an interest in token identities. He would no doubt reply that it is part of a mistaken conception of mind, but I would suggest a different answer: the apparent causal interaction between the mental and the physical.

Admitting causal relations here will require that we salvage some part of the idea of the mental as inner—and this, Collins will argue, takes us back to the mistaken theory. But it is a real question whether these relations *are* part of the theory or whether they may be invoked on the side of our pretheoretical thinking about mind. Collins suggests that it is positivist thinking about *explanation* which gives encouragement to identity theories of mind: if all explanation is causal explanation, then beliefs and desires are the obvious candidates to figure in the explanation of actions; these beliefs and desires are then taken to be inner realities that must be physical realities if they are to be causally effective in a physical world. In order to avoid this conception of the mind, we need to shift away from a causal account of reason giving. Collins suggests that teleological explanation should serve as the model for action explanation. This form of explanation "renounces reference to antecedent causes in favour of reference to the outcome" (144). He devotes the final two chapters of his book to a defense of the position that action explanation is not causal explanation, but he does not consider whether a causal condition is necessary for action. The argument in favor of the causal condition is as follows: a man may have a reason to act and he may act, without its being the case that that reason was the reason for that act. Collins may reply that such an argument relies on a prior acceptance of the misguided conception of mind. But one could rejoin by pointing out that it is our conception of action that demands the causal condition.

---

[1] *Naming and Necessity* (Cambridge: Harvard, 1980), p. 155.

The real strength of Collins's work is to be found in the wide perspective he offers. He forces us to re-examine so much of what has come to be taken for granted in some philosophical circles. It seems to me just right that what needs our attention these days is an examination of the conception of mind with which we are working. In this, Collins's work is exemplary.

ANITA AVRAMIDES

St. Hilda's College, Oxford University

THE JOURNAL OF PHILOSOPHY

VOLUME LXXXVIII, NO. 3, MARCH 1991

EPISTEMIC PATERNALISM: COMMUNICATION CONTROL IN LAW AND SOCIETY*

A popular principle in epistemology and the philosophy of science is the requirement of total evidence (RTE). A weak version of this principle may be stated as follows:

> (W-RTE) A cognitive agent *X* should always fix his beliefs or subjective probabilities in accordance with the total evidence in his possession at the time.[1]

This version says nothing about acquiring or collecting evidence, only about the use of evidence already in one's possession. A stronger version of RTE, however, addresses evidence gathering as well:

> (S-RTE) A cognitive agent *X* should collect and use all available evidence that can be collected and used (at negligible cost).[2]

This principle had best be understood in purely epistemic terms, rather than moral or legal terms. Some evidence collection might

---

* For valuable comments and advice, I am grateful to Holly Smith, Allen Buchanan, Henning Jensen, Joel Feinberg, Richard Lempert, and the members of the 1989 Dubrovnik conference on epistemology and philosophy of mind.

[1] See Rudolf Carnap, *Logical Foundations of Probability* (Chicago: University Press, 1950), p. 211; and Carl G. Hempel, "Inductive Inconsistencies," in *Aspects of Scientific Explanation* (New York: Free Press, 1965), pp. 64–7.

[2] See I. J. Good, "On the Principle of Total Evidence," in *Good Thinking* (Minneapolis: Minnesota UP, 1983); and Paul Horwich, *Probability and Evidence* (Cambridge: MIT, 1982), ch. 6. Good's argument for the strong version of RTE appeals to the criterion of maximizing expected utility, so it seems more pragmatic than purely epistemic. Horwich's argument, however, is purely epistemic. He shows that the *expected error* in one's probability judgment vis-à-vis some hypothesis H is minimized by the acquisition of any new evidence E (as long as $0 < \text{prob}(E) < 1$, and $\text{prob}(H/E) \neq \text{prob}(H/-E)$). I shall not discuss these arguments in detail because I am not so much interested in S-RTE as in the two-agent "control" version of RTE. In fact, I am not even primarily interested in the control version of RTE. I really use it as an expository heuristic to introduce the topic of epistemic paternalism, and the topic of communication control more generally.

0022-362X/91/8803/113–31

involve invasion of privacy or harmful experimentation on human subjects, which would be objectionable on moral and/or legal grounds. But if we abstract from these factors and restrict attention to epistemic considerations, the principle has initial intuitive appeal.

A plausible-seeming corollary, or extension, of S-RTE is a principle governing the practices of a second agent, *Y*, who is in a position to control the evidence made available to *X*. This interpersonal principle would say that *Y* should make available to *X* all evidence that is subject to his (*Y*'s) control. Of course, like S-RTE itself, the envisaged extension or corollary of S-RTE must be restricted to epistemic contexts or concerns. Thus, we might formulate the "control" version of RTE roughly as follows:

> (C-RTE) If agent *X* is going to make a doxastic decision concerning question *Q*, and agent *Y* has control over the evidence that is provided to *X*, then, from a purely epistemic point of view, *Y* should make available to *X* all of the evidence relevant to *Q* which is (at negligible cost) within *Y*'s control.

The restriction to the epistemic viewpoint is again important. In legal settings, for example, there are many nonepistemic reasons for refusing to provide relevant evidence to jurors. Available evidence may have been illegally obtained. Relevant evidence may be obtainable from the defendant, but the Fifth Amendment forbids his being compelled to testify against himself. Or the defendant may have testified elsewhere under a grant of limited immunity, providing that his testimony would not subsequently be used against him. In these cases, the judge (*Y*) is obliged not to provide the jurors (*X*) with all available evidence logically relevant to the question of guilt. These constraints, though, are not of an epistemic nature. Thus, the stated version of C-RTE may still be defensible. Indeed, anyone persuaded by John Stuart Mill's[3] famous thesis that, if we are interested in having truth prevail, we should allow all available arguments to be heard, should be attracted to C-RTE.

I shall argue, nonetheless, that C-RTE is unacceptable. More cautiously, I shall show that existing provisions and practices, both in the law and elsewhere in society, contravene C-RTE. Although I shall not defend each provision and practice in detail, many of them seem to be quite reasonable. This raises some interesting questions for a branch of epistemology that I have elsewhere called *social epistemics.*[4] The aim of this paper is to formulate these questions in a general way and to undertake a preliminary exploration of them.

---

[3] *On Liberty*, ch. 2.

[4] See *Epistemology and Cognition* (Cambridge: Harvard, 1986), pp. 1, 5–9, 136–8; "Foundations of Social Epistemics," *Synthese*, LXXIII, 1 (1987): 109–44;

I

The American legal system has a set of explicitly formulated rules of evidence for the federal courts. These and other trial-related procedures are substantially dedicated to the goal of getting the truth about the issues under litigation (in a criminal trial, the guilt or innocence of the accused). Since the aim of these rules is, accordingly, heavily or predominantly epistemic, they provide excellent examples for our scrutiny.

The Federal Rules of Evidence[5] resulted from proposals by an Advisory Committee appointed by the Supreme Court, which were passed into law (with some revisions) by Congress in 1975. Although the rules do not, in general, bind judges on the admission or exclusion of evidence, they provide guidelines that should be followed if decisions are not to be overturned. Rule 102 declares their purpose: "These rules shall be construed to secure fairness in administration, elimination of unjustifiable expense and delay, and promotion of growth and development of the law of evidence *to the end that the truth may be ascertained* and proceedings justly determined" (emphasis added). The truth goal is implicit in the declaration (Rule 602) that witnesses may only testify to a matter of which they have "personal" (especially perceptual) knowledge, and in the general exclusion of hearsay evidence (Rule 801). Such rules codify what McCormick calls a "most pervasive manifestation" of the common law, viz., insistence upon "the most reliable sources of information."[6] Reliance on reliable sources, clearly, is more likely to produce accurate (i.e., truthful) judgments. The admission of testimony by experts (Rule 702) also exemplifies a dedication to truth, since experts are people presumed to know relevant truths. The requirement of witnesses to make an oath or affirmation before testifying is still another device calculated to secure truthful evidence (Rule 603), as are statutory provisions of penalties for perjury. The entire procedure of cross-examination, which holds an exalted place in the Anglo-American trial system, is similarly rationalizable by reference to the truth goal. Since it is impossible to prevent false evidence from ever being introduced, at least there should be ample opportunity to contravert it, to reduce the likelihood of its being believed by the fact finder and thereby promoting a false verdict.

What is interesting in the present context is that the rules of

---

and "The Cognitive and Social Sides of Epistemology," in A. Fine and P. Machamer, eds., *PSA 1986*, II (East Lansing: Philosophy of Science Association, 1987), pp. 295–311.

[5] See *Federal Rules of Evidence for United States Courts and Magistrates* (St. Paul: West, 1989).

[6] In E. Cleary, ed., *McCormick on Evidence*, 3rd ed. (St. Paul: West, 1984), p. 23.

evidence frequently foster the exclusion of evidence from the jurors, in conflict with principle C-RTE. On the matter of admission or exclusion of evidence, one governing provision is that *relevant* evidence is generally admissible, but *irrelevant* evidence is not admissible (Rule 402).[7] In the matter of hearsay evidence, the basis for exclusion ostensibly falls under a different principle, not irrelevance but unreliability or doubtful veracity. Still other rules promote exclusion even when the evidence is both relevant and presumptively true. Evidence about the character of an accused is not admissible for the purpose of proving that he acted according to character on the occasion in question (Rule 404). In particular, evidence of previous *crimes* by the accused is not admissible to help prove that he committed the present crime. Also, evidence that the defendant initially entered a guilty plea on the present charge, and then withdrew it, is inadmissible (Rule 410). Here we have examples of a particularly interesting kind, in which judges are allowed or enjoined to exclude presumptively true and relevant evidence. Can these be rationalized on epistemic grounds?

The general rationale is given in Rule 403, which states in part: "Although relevant, evidence may be excluded if its probative value is substantially outweighed by the danger of unfair prejudice, confusion of the issues, or misleading the jury, ..." The phrase 'misleading the jury' apparently refers to the jury's being led into making an incorrect judgment, and the language of 'prejudice' and 'confusion' can also be readily interpreted this way. It appears, then, that the Federal Rules of Evidence provide an *epistemic* rationale for excluding various types of evidence from doxastic decision makers (jurors), contrary to what C-RTE enjoins.

II

A better understanding of the underlying rationale for the relevance rules can be gleaned from a theoretical reconstruction by a legal scholar, Richard Lempert.[8] Lempert offers several reconstructive suggestions, but I shall present just the one employing a Bayesian interpretation. The discussion assumes that the fact finder is a jury and the issue to be resolved is a defendant's guilt.

One form of Bayes' Theorem, suitable to the present case, is formulated as follows:

$$O(G/E) = \frac{P(E/G)}{P(E/\text{not-}G)} \cdot O(G)$$

---

[7] Relevant evidence is characterized as "evidence having any tendency to make the existence of any fact that is of consequence to the determination of the action more probable or less probable than it would be without the evidence" (Rule 401).

[8] "Modeling Relevance," *Michigan Law Review*, LXXV (1977): 1021–57.

This formula describes the way that a new item of evidence (*E*) should influence a rational agent's odds (*O*) that a defendant is guilty (*G*).[9] It says that the posterior odds of the defendant's guilt is equal to (1) the probability that the evidence would obtain if the defendant is in fact guilty, (2) divided by the probability that that evidence would obtain if the defendant is in fact not guilty, (3) multiplied by the prior odds on the defendant's guilt. The ratio of items (1) and (2) is conventionally called the *likelihood ratio*. On Lempert's reconstruction, a necessary and sufficient condition for evidence to be logically *irrelevant* to the defendant's guilt is that the likelihood ratio be 1:1 (or close to 1:1). This holds when the evidence in question would be just as likely to arise if the defendant were not guilty as if he were guilty, i.e., if $P(E/G) = P(E/\text{not-}G)$. Evidence of this sort has no "probative" value, that is, warrants no change in the posterior odds of guilt. Such evidence is unhelpful and therefore should not be admitted. Logically *relevant* evidence is evidence whose likelihood ratio departs (substantially) from 1:1.

Why should courts declare even relevant evidence inadmissible? Lempert expresses this in terms of the danger that the fact finder will misestimate the probabilities that comprise the likelihood ratio, and hence assign excessive (or insufficient) weight to the evidence. Overestimating the numerator or underestimating the denominator makes the conclusion sought by the proponent of the evidence appear more probable than it actually is; underestimating the numerator or overestimating the denominator has the opposite result. For example, suppose that in an assault case it can be shown that the defendant is a heroin addict, and also that one out of 500 criminal assailants are heroin addicts, whereas of the people who never engage in criminal assault only one in 1,000 are heroin addicts. Then knowledge that the defendant is an addict should result in a doubling of the prior odds that the defendant was the assailant. Suppose further that the fact finder (mistakenly) thinks that the probability that a nonassailant would be a heroin addict was one in 10,000 rather than one in 1,000. This misestimation of the denominator by a factor of ten leads to a twentyfold increase in the odds of guilt rather than a twofold increase. It is precisely this sort of danger which concerns the courts. Rule 403, excluding character evidence, is partly justified on this ground, says Lempert. Jurors are likely to magnify the import of character traits, or past criminal record;[10] and similarly for withdrawn guilty pleas. When courts or the codified rules speak of the problem as one of juror "prejudice," or when they speak of "confusing" or "misleading" the jury, this should be understood, according to Lempert, in terms of the jury's being prone to such misestimations.[11]

---

[9] I do not mean to endorse the Bayesian analysis for all purposes, including legal purposes. I am merely expositing Lempert's reconstruction. The same morals could well be drawn without commitment to Bayesianism. For an alternative account of reasoning principles for the law, see L. Jonathan Cohen, *The Probable and the Provable* (New York: Oxford, 1977) (which I also do not mean to endorse).

III

Whether or not Lempert's Bayesian reconstruction is wholly faithful to the intent of the rules, it is apparent that the framers of the rules, and judges themselves, often wish to *protect* jurors in their search for truth. If, in the framers' opinion, jurors are likely to be misled by a certain category of evidence, they are sometimes prepared to require or allow such evidence to be kept from the jurors. This is an example of what I shall call *epistemic paternalism*. The general idea is that the indicated rules of evidence are designed to protect jurors from their own "folly," just as parents might keep dangerous toys or other articles away from children, or might not expose them to certain facts. I do not wish to quibble here over precise definitions of the term 'paternalism'. My usage, however, has much in common with H. L. A. Hart's characterization of paternalism as "the protection of people from themselves," and with Joel Feinberg's emphasis on the analogy to parental relations with children.[12] Jurors may have flaws in their background beliefs, or in their ability to draw apt conclusions from evidence. If so, the courts are prepared to protect them against these information-processing deficiencies in order to get truthful judgments on the issues at hand. Admittedly, in the judicial case, the prime objects of "protection" are plausibly the parties to the litigation, or perhaps society at large, not the jurors. So the present use of the term 'paternalism' may extend a bit beyond its standard usage. The indicated parties are protected, however, by getting jurors to make accurate judgments. Protection of the *jurors'* epistemic ends therefore assumes derivative importance.

---

[10] In defending the rule against character evidence, the Advisory Committee for the Federal Rules of Evidence quotes the California Law Revision Commission with approval: "Character evidence is of slight probative value and may be very prejudicial. It tends to distract the trier of fact from the main question of what actually happened on the particular occasion. It subtly permits the trier of fact to reward the good man and to punish the bad man because of their respective characters despite what the evidence in the case shows actually happened" (*op. cit.*, p. 31).

[11] In Horwich's terms (see fn. 2 above), we might express the court's fear of jurors being misled by E as an assessment that the result of their getting E would be an *increase* in the probability of their making an error rather than a reduction. The crucial point, of course, is that the court substitutes its *own* expectation of juror error for that of the jurors themselves, who perhaps would welcome information about the accused's character, criminal record, or withdrawn guilty pleas as likely to reduce their probability of error.

[12] Hart, *Law, Liberty, and Morality* (New York: Vintage, 1966), p. 31; and Feinberg, *Harm to Self* (New York: Oxford, 1986), pp. 3–8. For other discussions of the definition of paternalism, see Rolf Sartorius, ed., *Paternalism* (Minneapolis: Minnesota UP, 1983); and Donald Van DeVeer, *Paternalistic Intervention* (Princeton: University Press, 1986), pp. 16–24.

In any event, I shall construe epistemic paternalism in a broad sense. I shall think of communication controllers as exercising epistemic paternalism whenever they interpose their own judgment rather than allow the audience to exercise theirs (all with an eye to the audience's epistemic prospects). Thus, the exclusion of evidence of doubtful veracity (e.g., hearsay evidence) also qualifies as epistemic paternalism. The courts apparently feel that jurors cannot be counted on to discount hearsay testimony adequately. So they substitute their own wisdom for that of the jurors. Similarly, when judges exclude evidence as irrelevant, they must use *their* assessment of whether the evidence affects the probabilities of the propositions in question (the jurors' assessments might have been different). When these categories of exclusion are added, it is clear that courts engage in a substantial amount of epistemic paternalism.

Is such paternalism really warranted? Are these rules *good* rules from an epistemic point of view? This is open to dispute. Lempert contends that evidence should not be excluded when the probative value is substantial, i.e., when the likelihood ratio deviates markedly from 1:1. The preferred solution, he says, is to provide the jury with the information needed to assess accurately the probative value of the offered evidence.

There are general problems with this solution, though. For one thing (as Lempert notes), there may not be hard data that indicate just how heavily a piece of evidence should be weighted. There may be no well-researched (base rate) "facts" to give to the jury. Nonetheless, people with judicial experience may have good reasons to suspect that juries would exaggerate the import of that type of evidence. Second, there is a question of whether statistical information, even if it were available, would psychologically displace or override jurors' prior prejudices.[13] Third, there is a question of whether *all* the relevant considerations that bear on the import of the evidence should be given to the jurors. Should they be presented with the Bayesian framework in terms of which the evidence may be interpreted? Should they be informed of theoretical disputes over the correctness of Bayesianism, or the potential for its misapplication? Much of this material would surely confuse most jurors. The dangers of introducing mathematical techniques into the fact-finding process have been emphasized by Laurence Tribe.[14]

---

[13] Studies by Lee Ross and colleagues give evidence of a "perseveration" phenomenon, in which previously acquired beliefs are not eradicated even by new evidence showing that the original basis of those beliefs was ill-founded. See Richard Nisbett and Lee Ross, *Human Inference: Strategies and Shortcomings of Social Judgment* (Englewood Cliffs, NJ: Prentice-Hall, 1980), ch. 8. I have suggested more generally that beliefs can never be "erased," but at best only "overridden." This allows for the presence of *credal residues*. See *Epistemology and Cognition*, pp. 223–6.

I am inclined to think that *some* paternalism is appropriate in this arena, although I shall not take a firm stand on specific policies. What I want to do is identify the questions clearly, and put them in the framework of a wider set of questions. What we have here is a set of rules or practices whose adoption has an impact on the truth values of the doxastic decisions that cognizers make. One question is: For each such rule, how good is its impact from a *veritistic* point of view, that is, in terms of the likelihood of getting truth and avoiding error? Would alternative rules or practices have better veritistic properties?

This type of question falls under what I call *social epistemics*. Epistemics generally is my (partly "reforming") conception of epistemology, with three main divisions: primary individual epistemics, secondary individual epistemics, and social epistemics. All three divisions seek to make evaluations in terms of veritistic ends, i.e., in terms of the effect on people's getting truths and avoiding errors. The objects or targets of evaluation, however, differ from division to division. Primary epistemics studies the veritistic properties of basic psychological processes. Secondary epistemics assesses the veritistic properties of learnable problem-solving methods, such as mathematical proof techniques, carbon-dating procedures, and the like. Social epistemics studies the veritistic properties of social practices, or institutional rules that directly or indirectly govern communication and doxastic decision. Judicial rules of evidence clearly fall in this last category.[15] The focus of the present paper is the class of institutional rules or practices that have a flavor of epistemic paternalism, e.g., those which weed out some possible communications that might be directed to a cognitive agent, by appeal to that agent's own veritistic ends. Are such rules and practices defensible, and, if so, which ones?

---

[14] "Trial by Mathematics: Precision and Ritual in the Legal Process," *Harvard Law Review*, LXXXIV, 6 (1971): 1329–93.

[15] For further discussion of the divisions of epistemics, and the rationale for them, see "Foundations of Social Epistemics" and "The Cognitive and Social Sides of Epistemology." For another treatment of judicial rules of evidence in the context of social epistemics, see Keith Burgess-Jackson, "An Epistemic Approach to Legal Relevance," *St. Mary's Law Journal*, XVIII, 2 (1986): 463–80. I thank Burgess-Jackson for helping draw my attention to relevance rules.

Although social epistemics could content itself with the separate evaluation of individual practices, it seems more advisable to address whole classes of practices. What one wants is a general assessment of paternalistic control practices, first in terms of purely epistemic considerations and second on more inclusive grounds. Before turning to general principles, however, let us examine some additional cases of paternalistic practices.

IV

Start with curriculum selection in education, especially in primary and secondary schools. School personnel at various levels—boards of education, principals, and teachers—select curricular materials in the form of textbooks, course syllabi, and so forth. (When teachers do it, they assume something of a dual role: both controller and speaker.) In the nature of this business, some points of view and supporting argumentation are left out. Students are not exposed to all possible ideas on a given subject. Is this objectionable? Set aside highly publicized examples of book banning based on obscenity, profanity, or offensiveness. Think instead of simply ignoring opinions that have (or once had) their exponents, but are regarded by current authorities as palpably false or indefensible. Mathematics classes do not present alternative (crackpot) mathematics. Science classes do not (often) present the flat-earth viewpoint, Ptolemaic astronomy, or astrology. Schools rarely if ever invite Jeane Dixon or her ilk to give guest lectures, or recount as serious alternatives the theories of Velikovsky. Classes in health education do not give "equal time" to drug pushers to defend the safety of drug use, or to quacks to present and defend their cures. These omissions probably have veritistically good consequences. Of course, it might also be appropriate to have classes sample and compare "bad" and "bogus" science along with the good.[16] But this kind of course might be advantageous only for high-school students, not for younger pupils, who would only be confused.

The most public controversy surrounds the question of "equal time" for creationism in biology classes. Here, too, paternalism seems to be warranted on epistemic grounds. Experts on science should be allowed to decide that creationism is not a scientifically viable or serious contender, and hence should not be taught in the classroom.[17] What about the teaching of creationism in nonscience classes? Is it legitimate for a local community to require this in their schools? Here extra-epistemic issues intrude. Creationism is presumably a religious doctrine, so that constitutional issues arise about its being taught in public schools. What about the putative right of parents to control their children's education? If such an alleged right is supposed to rest on purely epistemic considerations, it would be very hard to defend. If it is based on nonepistemic considerations of an ethical or sociopolitical nature, it goes beyond the main focus of our discussion, although I shall briefly return to it in section VII.[18]

---

[16] The phraseology is borrowed from the title of Martin Gardner's book, *Science: Good, Bad, and Bogus* (Buffalo: Prometheus, 1981).

[17] This is not to dispute Paul Feyerabend's point [*Science in a Free Society* (London: Verso, 1978)] that the expertise of science and scientists must ultimately be judged by laymen. The thesis that laymen can assess the credentials of putative experts is discussed below in sect. VIII.

Turn next to the sphere of commercial advertising, where the Federal Trade Commission has authority to take actions against false or deceptive advertising. This exemplifies epistemic paternalism because it prevents some potential messages from being conveyed to an audience, and it does so with veritistic ends in mind: to keep them from believing untruths about commercial products. The FTC has four types of remedies for inhibiting false and deceptive advertisements, or even requiring truthful disclosures that the "speakers" would not independently make.[19] First, it can issue *cease and desist* orders. Second, it can seek to obtain *affirmative disclosure,* forcing a manufacturer to disclose by mark or label material facts concerning the merchandise. The most familiar such example is the requirement of warning labels on cigarette packages (although this is under the jurisdiction of the Food and Drug Administration, rather than the FTC). A third remedy is *corrective advertising*. When consumer misconceptions about a product have resulted from past advertising, the FTC has sometimes forced the manufacturer to devote 25% of its advertising budget for one year to corrections of past inaccuracies. In this arena, then, communication control sometimes takes a stronger form than the one considered thus far: not excluding messages, but mandating messages of a specified kind or content. A fourth remedy is the FTC's advertisement *substantiation* program. This forbids manufacturers to make certain advertising claims unless they have done competent scientific tests that substantiate their claims. This piece of epistemic paternalism has been rationalized by appeal to the relatively poor opportunity of consumers to obtain accurate information themselves.[20]

---

[18] On issues of educational philosophy and the control of educational policy, see Amy Gutmann, *Democratic Education* (Princeton: University Press, 1987).

[19] On these topics see Donald Gillmor and Jerome Barron, *Mass Communication Law,* 4th ed. (St. Paul: West, 1984), ch. 8; and Daniel Rohrer, *Mass Media, Free Speech, and Advertising* (Dubuque, IA: Kendall/Hunt, 1979), ch. 3.

[20] FTC Chairman Miles Kirkpatrick wrote: "The manufacturer has the ability, the knowhow, the equipment, the time and the resources to undertake such information by testing or otherwise—the consumer usually does not" (quoted in Gillmor and Barron, pp. 618–9).

Another class of examples involves television and radio news. Generally speaking, American network news broadcasts offer a relatively limited variety of interpretations of each news event; nor do they detail the evidence for each of the possible interpretations. This is partly due, no doubt, to severe time constraints. But there is also the deliberate attempt to simplify, to make the news understandable and digestible to a large audience. Certainly this must be included under the rubric of epistemic paternalism. There is room for debate over how acceptable this practice is, and doubtless different programs and newscasters deserve different ratings on this dimension. But it is hard to deny the need for some degree of simplification, especially on technical topics. If this is at all legitimate, some degree of epistemic paternalism again seems inevitable and unobjectionable.

Simplification involves the omission of some truths. So paternalism in this form reduces the number of truths an audience has an opportunity to acquire. Nonetheless, the trade off of error avoidance, or confusion avoidance, may compensate for this loss (depending partly on how different segments of the audience are affected).

A further element in the trade off is that simplification can increase the audience size. This may mean that more true beliefs (i.e., belief tokens) are acquired through the simplified set of messages than would otherwise be the case. Although members of the media may care about this chiefly because of Nielsen ratings, and their consequent attractiveness to advertisers, the number of people acquiring true beliefs is a genuinely epistemic value, which I elsewhere call *fecundity*.[21]

The epistemic consequences of an institutional policy are often difficult to anticipate. The Federal Communication Commission's "Fairness Doctrine" is a good case in point. Since holders of its broadcast licenses are supposed to operate in the public interest, the FCC has required licensees to devote a reasonable amount of time to issues of public importance. Moreover, if a station presents one side of a controversial issue of public importance, it must afford reasonable opportunity for presentation of contrasting views. This policy was intended to promote veritistic outcomes by exposing the public to a wide diversity of opinion. It has been argued, however, that its actual result is opposite to its intent. The threat of incurring FCC suits has allegedly discouraged stations from airing controversial material. This "chilling" effect has tended to guarantee blandness, and to retard rather than enhance a free marketplace of ideas. This is a good example of how a policy's epistemic consequences (historical or prospective) can be a complex empirical question. This is why empirical sciences have a role to play in social epistemics.

---

[21] See "Foundations of Social Epistemics" and "The Cognitive and Social Sides of Epistemology."

V

Let us now inquire into the general circumstances in which epistemic paternalism, and other communication control policies, have good or bad epistemic consequences, i.e., good *veritistic* outcomes. A number of variables are relevant, especially: (1) the characteristics of the *controller* (or "gatekeeper"), (2) the characteristics of the *speakers* who wish to send messages via the communication channel, (3) the controller's *criterion of selection* among speakers or messages, (4) the characteristics of the *audience*, and (5) the availability of *alternate channels* that address the same topic.

Epistemic outcomes obviously depend heavily on the (epistemically) good judgment of the message controller. If the controller accurately distinguishes true and false claims, and/or true and false pieces of evidence for claims (and their relevance), then the choice of included and excluded messages may well promote truth acquisition on the part of the audience. This brings us immediately to variable (3): the criterion of selection. Does the controller select among candidate messages by reference to (A) their specific content, or (B) the characteristics of the speaker (or a combination of the two)? If the messages directly attempt to answer primary questions in the domain, and the controller himself has expertise on these questions, then a selection based on message content may positively contribute to veritistic outcomes. The controller may not purport to know answers to the primary questions in the domain, however. He may only select messages on the basis of prospective speakers' "credentials." Was the speaker an eyewitness to the putative events he wants to describe, or does he have some other access to, or authority about, the facts in question? If the controller is to make veritistically good selections, he must at least have expertise on the question of speaker credentials. (We might call this *secondary*, as opposed to *primary*, expertise.) This is what judges or rule framers claim to have in the field of legal evidence.

Selection criteria do not always involve purely veritistic considerations. A network may shy away from certain political interpretations because of prospective sponsor or audience disapproval. Since this is not a truth-oriented selection rationale, it raises doubts about the epistemic optimality of the control process.

Speaker characteristics are equally critical to (audience) veritistic outcomes. We have already mentioned the "access" or authority of speakers on the matters whereof they speak. Of almost comparable importance are the speakers' motivational properties. Do prospective speakers have an incentive for accuracy, or do they have incentives for deception and misrepresentation, as in the case of manufacturers, partisan witnesses, and perhaps even research scientists? If there are incentives for deception, does the *controller* know or suspect them, and does he use them in his selection policy?

Turning to audience characteristics, how informed, sophisticated, rational, etc., are the audience members who are likely to receive the channel's messages? How good are they at weighing competing views or complex chains of evidence? How good are they at assessing the reliability of various speakers or sources of evidence? Clearly, all these traits are crucial to veritistic outcomes. Furthermore, to the extent that the controller takes account of audience characteristics in selecting messages, the controller's knowledge (or true belief) about their characteristics is critical. If the controller has an accurate assessment of the audience's cognitive limits or shortcomings, a paternalistic choice of messages to include (or to *mandate*) may avert potential audience error. Or his policy of selection may avert potential *confusion,* which can lead to doxastic indecision or simple incomprehension of the truth (either of which is a comparatively bad veritistic outcome).

Finally, the availability of alternate communication channels can influence veritistic outcomes. If one channel deploys restrictive selection policies, that does not necessarily mean that the audience is confined to those messages. There may be other potential sources of information.

These are some of the salient factors which bear on the epistemic success of various alternative institutions of communication control. They are the sorts of factors to which social epistemics should appeal in evaluating existing or prospective institutions. As the discussion indicates, institutions or policies of these sorts cannot be rated in the *abstract*, apart from the properties of the individuals who occupy, or are likely to occupy, the different institutional roles. The same institution or policy might work well if the controller has considerable expertise, but poorly if he does not. In any case, when the cited variables or parameters take on appropriate values, epistemic paternalism will be justified. My previous examples are, quite plausibly, instances of this sort. So C-RTE must be rejected as a universal principle.

VI

I have been equating epistemically valuable outcomes with true belief and error avoidance. Are there additional epistemic values that this ignores? What about Mill's value of a "clearer perception and livelier impression of truth," produced by its collision with error (*op. cit.*, ch. 2)? And what about the skills to be learned from the process of wrestling with competing doctrines? I suspect that the livelier impression of truth, of which Mill speaks, can be cashed out in terms of a grasp of a larger number of related truths. One learns not simply that answer A_1 to question Q is correct, but that certain other answers, A_2, A_3, etc., have been offered to Q, what the arguments for and against these alternative answers are, and why these alternatives are mistaken. Learning these things doubtless has epistemic value, but it is not an additional *kind* of value. Similarly, the value of argumentative skills can be cashed out in terms of their instrumental value for discovering further truths on one's own, when no instructor is available.

Thomas Scanlon[22] has expressed doubts about epistemic protectionism by appeal to the value of autonomy, which consists in a person seeing himself as sovereign in deciding what to believe and in weighing reasons for action. Is autonomy an epistemic value of which a person is deprived by epistemic paternalism? If sovereignty is a matter of reserving the final doxastic choice to oneself, as some of Scanlon's passages suggest, this is not compromised by epistemic paternalism. The juror who did not hear certain excluded evidence still has to decide whether to believe what was presented. Perhaps the point, however, is that, whenever a cognitive agent is deprived of some doxastic alternative, or some evidence relevant to that alternative, some degree of sovereignty is reduced. This may be granted, but does this reduction constitute a diminution in *epistemic* value (especially *intrinsic* epistemic value)? I doubt it, though the matter is not wholly clear. An analogous case involves "practical" value, and the range of alternatives actually scanned by a deliberating agent. Suppose agent S wishes to select a plan of action designed to achieve goal G. If an assistant fails to call S's attention to some unnoticed alternative plan which is in fact worthless (it would not achieve G at all, or would be much too costly), S's planning sovereignty has been reduced; he loses an opportunity to deliberate about the unpresented alternative. But is this a loss in "practical" value? Again, I am doubtful.

Even if we concede a measure of (intrinsic) epistemic value to autonomy, a value compromised in epistemic paternalism, it is questionable whether the loss always outweighs the gains. As John Hardwig[23] has stressed, we live in an epistemically complex world, where each of us cannot reasonably hope to assess all evidence for all theses personally. We often have to depend on the authority of others.

---

[22] "A Theory of Freedom of Expression," *Philosophy and Public Affairs*, I, 2 (1972): 204–26.

[23] "Epistemic Dependence," this JOURNAL, LXXXII, 7 (July 1985): 335–49.

Given this situation, it seems likely that epistemic paternalism will frequently be necessary, and sometimes epistemically desirable.

VII

The burden of the preceding sections is that epistemic paternalism can sometimes be warranted on *epistemic* grounds. Extra-epistemic considerations, however, should unquestionably enter into overall assessments of communication control policy. What are these other considerations?

Let me briefly suggest five additional factors that ought to be considered. First is the *practical significance* factor. Some communications induce audience beliefs that are likely to produce immediate actions with serious outcomes, including harms either to themselves or to other agents. (This is similar to the "clear and present danger" idea.) Direct and serious outcomes are salient in a trial, where juror beliefs dictate court sanctions that directly affect the defendant's welfare (and the welfare, one might say, of society).[24] Similarly, consumer beliefs about commercial products standardly lead to consumption choices, which may directly affect health or safety. These are the cases where epistemic paternalism seems most appropriate. Consumers' welfare, for example, seems to outweigh any presumed right of manufacturers to send any message they like.[25]

A second factor is the *power of the controlling agent*. Epistemic paternalism on the part of isolated individuals is quite a different matter from paternalism exercised by the state, or any other powerful organ of society. There are historical reasons for being very cautious about state control of information.

Closely related to this is the third factor: the *scope* of control. The greater the breadth, the greater the dangers of communicational restriction. This was already mentioned in section V, but is worth repeating here. The exclusion of certain viewpoints from school curricula seems more defensible when it is recalled that students have other communication channels available. They can be taught things (e.g., religion) outside of school. In the case of broadcast news, the public is free to select among numerous stations, and to read newspapers, books, and magazines as well. Thus, the extent of control of any single news outlet is *partial* (and small). The chief objection to state censorship, by contrast, is that it is *total*. This is a very different kettle of fish. My condoning of local epistemic paternalisms should not be taken as an endorsement of global state censorship, especially on topics like news, politics, and public affairs.

---

[24] Of course, where the defendant is in fact guilty, his welfare is not served by finding the truth. Presumably, however, society's utility is increased by accurate verdicts.

[25] There is ongoing debate, however, over the extent to which commercial speech is protected under First Amendment guarantees of freedom of speech. Not everyone agrees that regulation of advertising is legally defensible, or a good idea. See, for example, R. H. Coase, "Advertising and Free Speech," *Journal of Legal Studies*, VI, 1 (1977): 1-34.

This brings me to a fourth factor: the *topic* of the messages being controlled. It has been contended—quite plausibly, in my opinion—that an unregulated marketplace of ideas seems essential for political speech but not commercial speech. Thus, policies should differ for different subjects. I mention this here under nonepistemic factors; but actually it *may* be explainable by reference to epistemic ones. Perhaps we feel comfortable with a mathematics or science authority omitting nonstandard treatments from the curriculum because we acknowledge their expertise in their subjects. By contrast, in matters of (normative) politics and religion, we have no analogous conviction—as a society, at any rate—that there are any experts, and certainly no societal consensus on who the experts are (if any).

A fifth factor (or group of factors) concerns the putative *rights* of concerned agents: a speaker's right to freedom of speech, a communication channel owner's right to run his own station or publication, a parent's right or a community's right to predispose their children toward certain values or heritages, and so on. As indicated, some of these (putative) rights may be rationalized by epistemic factors, but others may be independent. In each case, moreover, the right may not be unconditional, but must be balanced by other rights or interests.[26]

How these nonepistemic factors should be balanced along with the (purely) epistemic ones is a complex matter, and I have no formula to offer. The variability of situations along these dimensions, however, makes it reasonable to expect epistemic paternalism to be appropriate (all things considered) in some contexts though inappropriate in many others.

VIII

Since expertise is a prime epistemic factor in the defense of epistemic paternalism, more should be said about it. To justify any particular instance of such paternalism, involving a particular controller, we must have grounds for taking that agent to be an expert. But how, it may be asked, are people supposed to *tell* who (if anyone) has expertise? In particular, if (primary) expertise is defined in terms of believing domain-relevant truths, how can people lacking expertise —novices—identify an expert? Furthermore, when different people hold conflicting views and each claims to be an expert, how can the issue be resolved to their mutual satisfaction?

---

[26] Scanlon has emphasized that issues in the freedom of expression hinge on the interests of three parties: participants (speakers), audiences, and bystanders. See "Freedom of Expression and Categories of Expression," reprinted in David Copp and Susan Wendell, eds., *Pornography and Censorship* (Buffalo: Prometheus, 1983). In cases of communication control, however, the interested parties are even more numerous. And in the case of children, some sort of special status (not merely "bystander") seems appropriate for parents and community.

The problem of *expert identifiability* involves large epistemological issues that can only be addressed here partially, not fully. First, let us define an expert as someone who either (1) *has* true answers to core questions in the domain (i.e., believes or assigns high probability to these answers), or (2) has the *capacity* (usually through the possession of learned methods) to acquire true answers to core questions when they arise. Degree of expertise, then, is primarily a function of the candidate's question-answering *power,* i.e., how often he can form a belief in a correct answer as opposed to having either no opinion or an incorrect opinion.[27] Expertise can also be understood in either a *comparative* or an *absolute* sense. Someone is comparatively expert if his question-answering power ranks high compared with others; absolutely expert if his power ranks high in absolute terms.

The crucial requirement for expert identifiability is that, once candidate experts have produced their answers to target questions, there are "truth-revealing" situations in which novices and rivals can recognize which answers (if any) are correct. With enough such cases, the genuine experts can be identified. In fact, there are many domains in which such truth-revealing situations exist. Let me catalogue four homely domains of this sort: (1) *prediction* domains, (2) *factual-record* domains, (3) *repair* domains, and (4) *design* domains.

Prediction domains are ones in which the core questions are questions about future events, e.g., the weather, the economy, the stock market, or an upcoming election. When the events actually occur, the correct answers to the questions are revealed.

Factual-record domains can be illustrated with baseball and opera. An expert in these fields can correctly answer (from memory) questions about baseball or opera facts, e.g., who stole the most bases in 1973, which arias are from which operas, and so forth. Here there are undisputed records to which novices and rivals can appeal to verify correctness.

---

[27] The epistemic importance of question-answering power is emphasized in my *Epistemology and Cognition,* ch. 6. A further dimension in the definition of expertise should probably be added: the ability to recognize *wrong* answers as wrong (or assign them a low probability). Even if nobody has a correct answer to question *Q,* someone who knows (or truly believes) that each of the available answers is wrong is better off, epistemically, than someone who mistakenly believes (or assigns high probability to) wrong answers.

System-repair cases are more interesting. Whether a system (e.g., an air conditioner, an automobile, or a human body) is functioning properly or malfunctioning is frequently an uncontested, novice-detectable affair. Suppose system *S* is malfunctioning, so the question arises: "What techniques, steps, operations, or therapies would be followed by *S* returning to proper functioning?" A novice will typically have no answer at all. An expert, by contrast, can often answer correctly. The expert electrician, mechanic, or physician commonly proceeds as follows. First, he tries to find a system component that is causing the malfunction, using gauges, instruments, or other diagnostic techniques. Second, having diagnosed the cause, he decides what part replacements, alterations, adjustments, or medications would correct the malfunction. He thereby arrives at an answer to the target question. Can the truth of his answer be checked? Often it can. If and when the prescribed measures are executed, it will be novice-detectable whether the system resumes normal operation.

The fourth category of domains—design domains—is a fairly heterogeneous lot. Here the core question is: "What specific design or performance features would produce a certain desired effect, or meet certain standards or specifications?" This is the sort of question that sundry artists, designers, product engineers, or even mathematicians might be asked. The engineer might be asked for a blueprint of a dam that will meet certain specifications. A mathematician might be asked for a sequence of steps that would comprise a proof of a given theorem (according to understood constraints on the nature of proof).[28] In each of these cases, novices could not answer such questions correctly very often; but an expert could. (At least he could indicate correctly which answers are erroneous.) Is the correctness of an answer novice-detectable? Often the answer is 'yes'. Unable to produce a proof himself, the novice mathematician may nonetheless be able to verify whether someone else's proposal constitutes a proof. Whereas it may be difficult to design a theater set or choreograph a sequence that produces a desired effect, it may be palpable and uncontroversial that a certain design or sequence *does* (or does not) achieve that effect.

The moral is that, in many familiar cases, it is relatively easy, even for novices or rivals, to identify expertise. Needless to say, not all cases are straightforward. There is no necessary 'paradox', however, or impossibility in principle, of expert identification. There is no theoretical objection, then, to invoking the notion of expertise, construed veritistically, in the evaluation of communication practices.

---

[28] Even certain questions of theoretical science might be construed in this fashion, namely, "What is a theory in domain *D* that meets all the known empirical and conceptual constraints?"

IX

Let me return to the theme of epistemic paternalism, first illustrated with the rejection of C-RTE. To the extent that traditional epistemology has been attentive to the existence of multiple agents, it has generally assumed an idealized setting, in which all agents have the same cognitive resources, skills, and opportunities, where there are no time constraints, and so forth. In that sort of setting, a communication control principle like C-RTE might make sense. But in settings marked by different levels of expertise, by different opportunities for information gathering, by different levels of cognitive maturity and training, and by severe time constraints, idealized principles of communication do not plausibly apply. A social epistemology for the real world needs to take these constraints into account.[29]

The very question of communication control policy is relatively neglected in philosophy,[30] perhaps because the topic has been so thoroughly dominated by the Millian dedication to a free market of ideas. Interestingly, though, organized science and scholarship are very far from laissez-faire marketplaces. On the contrary, professional journals rely heavily on (putatively) expert referees and editors to weed out inferior contributions. Only offerings that are judged methodologically sound, well-informed, and possibly in the direction of truth are accepted for publication (or, in a related arena, given research funding). Thus, even in the purely intellectual sphere, laissez faire is not the de facto policy. More precisely, there is a complex layering of institutions. In free societies, governments pursue the laissez-faire policy of allowing private information channels to proliferate. But these information channels may themselves be highly restrictive in the messages they transmit. It could be argued that successful pursuit of epistemic ends depends not only on "deregulation" at the highest level, but on wise regulation at lower levels. This problem needs a more subtle and systematic examination than it has hitherto received. The present paper is intended to help motivate such an examination.

ALVIN I. GOLDMAN

University of Arizona

---

[29] I stress a similar theme of psychological realism in *Epistemology and Cognition*. Issues raised for moral theory by cognitive limits are stressed by Holly M. Smith in "Making Moral Decisions," *Nous,* XXI (1988): 89–108, and "Two-Tier Moral Codes," *Social Philosophy and Policy,* VII, 1 (1989): 112–32.

[30] One author who stresses the role of communication in a theory of rationality, however, is Jürgen Habermas, *The Theory of Communicative Action, vol. 1, Reason and the Rationalization of Society,* Thomas McCarthy, trans. (Boston: Beacon, 1981).

EPISTEMOLOGY ON HOLIDAY*

SOME years ago, an acquaintance complained to me about her first—and only—philosophy course. Her instructor had devoted a substantial portion of the course to the question of what—if anything—justified his students' belief that there was a lectern at the front of the classroom. To my acquaintance, the entire exercise had seemed a sham. It was obvious to her that there was a lectern in the front of the room and it seemed, as far as she could tell, that it was equally obvious to everyone else in the classroom, the instructor included. And it seemed to her intellectually dishonest of the instructor, and those students whose interest he had managed to engage, to pretend to throw into question the propriety of a belief when, in fact, the truth of that belief was entirely evident to them.

I had heard the complaint before—it was a complaint I had received from some of the students to whom I had taught epistemology. And it was a complaint to which I had a ready answer. The answer was Roderick Chisholm's.

I

In "Theory of Knowledge in America,"[1] and the subsequent two editions of his monograph *Theory of Knowledge*,[2] Chisholm contends that those who put forward such complaints are guilty of having misunderstood the question that is being asked them. Understood as a challenge to the propriety of the belief, the question 'What justification do you have for your belief that there is a lectern at the front of the classroom?' is indeed misguided. But, Chisholm continues, that is not how the question should be understood. We choose to focus upon the belief that there is a lectern at the front of the classroom precisely because we are antecedently convinced that we *are* justified in believing it is true. Our question about that belief, far from meaning to cast doubt upon its propriety, actually presupposes that our belief is justified. The question is simply intended to elicit the nature of that justification.

---

* I would like to thank Richard Foley, Miriam Solomon, William Wainwright, Joan Weiner, and the More or Less Exact Philosophy Discussion Group at the University of Cincinnati for helpful criticisms of earlier versions of this paper.

[1] Repr. in R. M. Chisholm, *The Foundations of Knowing* (Minneapolis: Minnesota UP, 1982), pp. 109–93, esp. pp. 130–1; orig. in *Philosophy, The Princeton Studies: Humanistic Scholarship in America* (Englewood Cliffs, NJ: Prentice-Hall, 1964).

[2] Englewood Cliffs, NJ: Prentice-Hall, 1966, pp. 24–5; in 2nd ed. (1977), pp. 16–8.

0022-362X/91/8803/132–54

Chisholm is quite explicit—and consistent over the course of the eleven years or so between the first and third of these writings—about how he conceives of this process of elicitation, what he calls the process of "investigating the theory of evidence from a philosophical—or Socratic—point of view."[3] I quote from the second edition of *Theory of Knowledge:*

> In order to formulate, or make explicit, our rules of evidence, we will do well to proceed as we do in logic, when formulating the rules of inference, or in moral philosophy, when formulating the rules of action. We suppose that we have at our disposal certain instances which the rules should countenance or permit and other instances that the rules should reject or forbid; and we suppose that by investigating these instances we can formulate criteria which any instance must satisfy if it is to be accepted or permitted, as well as criteria which any instance must satisfy if it is to be rejected or forbidden. To obtain the instances we need if we are to formulate rules of evidence, we may proceed in the following way.
>
> We consider things that we know to be true, or think that we know to be true, or certain things which, upon reflection, we would be willing to call evident. With respect to each of these, we then try to formulate a reasonable answer to the question, "What justification do you have for thinking you know this thing to be true?" or "What justification do you have for counting this thing as something that is evident?" In beginning with what we think we know to be true, or with what, after reflection, we would be willing to count as being evident, we are assuming that the truth we are seeking is "already implicit in the mind which seeks it, and needs only to be elicited and brought to clear reflection."[4]

Thus, for Chisholm, there is nothing intellectually unrespectable about devoting a large portion of an introductory course in philosophy—or, indeed, as Chisholm himself has done, a large portion of a philosophical career—to seeking an answer to a question like 'What, if anything, justifies your belief that there is a lectern at the front of this classroom?' provided one understands this question is "Socratic and therefore not at all of the type that one ordinarily asks."[5]

---

[3] *Theory of Knowledge*, 2nd ed., p. 16. The third edition of *Theory of Knowledge* (1989) is mute on the nature of the process of elicitation. See fn. 7 below.

[4] *Theory of Knowledge*, 2nd ed., p. 17. Chisholm is quoting, at the end of the passage, from C. I. Lewis, *Mind and the World Order* (New York: Scribner, 1929), p. 19.

[5] *Theory of Knowledge*, p. 25; 2nd ed., p. 18. Some may object that Chisholm's question is not really that similar to the question that provoked my acquaintance's complaint: (i) 'What if anything justifies your belief that there is a lectern at the front of the classroom?' The foregoing paragraphs indicate Chisholm would instead be inclined to ask: (ii) 'What justification do you have for counting it as evident—for

II

I do not recall whether, or in what detail, I presented my acquaintance with this answer to her complaint. Nor do I recall whether the answer satisfied her. All I know is that the answer no longer satisfies me. It is not Chisholm's assessment of the import of the question as ordinarily understood that rankles. In ordinary contexts, requests for the justification for a belief are indeed challenges to the propriety of the belief. To accept such a challenge and fail to meet it requires the admission that the belief in question is unfounded and ought to be abandoned. As a consequence, we tend to be rather careful in deciding which challenges we are going to treat as legitimate—which challenges we are going to feel bound to accept. The decisions we make define our methodology. They determine the claims for which we will feel bound to produce justification, the claims for which we will not, and they determine what sort of justification we will feel bound to produce.

This being so, it is obvious why Chisholm should not want to be understood as urging that we enter philosophy classrooms and ask, as if in an ordinary context, "What justifies your belief that there is a lectern at the front of the classroom?" He does not mean for us to cast aspersion on the propriety of our students' belief. He does not mean for us to challenge the soundness of that methodology which sanctions their belief. It is not that our students provide poor targets at which to aim methodological advice. They come to us with many

---

thinking that you know—that there is a lectern at the front of the classroom?' And questions (i) and (ii) may not seem to be asking the same thing.

But, as William Alston has persuasively argued [in "Some Remarks on Chisholm's Epistemology," *Nous,* XIV, 4 (November 1980): 565–86, esp. 580–1], much of what Chisholm has to say about what justifies one's counting propositions as evident—as opposed to what justifies one's believing propositions to be true—is at best poorly motivated, at worst a product of confusion, and, in any event, an unnecessary embarrassment to the theory of justification Chisholm actually offers. This theory, Alston concludes, is much more successfully viewed as a response to something like the following question: (iii) 'What justification do you have for thinking that there is a lectern at the front of the classroom?'

If Alston is right (as I think he is) and Chisholm is thus most charitably read as attempting to provide us a way to answer question (iii), then there is no obstacle to reading him as I have done: as attempting to provide us a way to answer to question (i). It simply needs to be added that Chisholm is to be distinguished from other epistemologists who offer answers to question (i): he insists that only such justification that a person *has*—only such justification that a person "can know just by reflecting upon his own state of mind" (*Theory of Knowledge*, 3rd ed., p. 7)—is capable of justifying that person's beliefs.

unreasoned prejudices, full of beliefs which are open to legitimate challenge and which deserve to be withheld until such time as the challenge is met. But surely their belief in the existence and location of the lectern is not among them. And just as surely the challenge that would, if accepted, have them abandon that belief until or unless they produce an appropriate argument on its behalf is a challenge that they can safely decline without fear of sinning against the strictures of good methodology.

What bothers me is the way Chisholm would reconstrue the question. To see why I am bothered, imagine that you are sitting in a doctor's crowded waiting room. The novel you have brought to read while you wait has just become interesting when the man sitting three seats to your left rises, points to you, and declares, in a voice loud enough for all to hear, "This person is a traitor to our country!" Needless to say, you are taken aback. "What, if anything, justifies your belief that I am a traitor?" you ask the man. Your accuser is calm. He explains that, although he recognizes that in accusing you of treason he has made the sort of claim that may well require justification, he cannot at present produce the justification for you. Nonetheless, he continues, he is confident that he is indeed justified in his belief that you are a traitor and, thus, he is confident that, with enough time and effort, an appropriate justification for this belief can be elicited from him.

I suspect that I know what you will want to say to your accuser. You will want to say that it is irresponsible to make a claim—and, more to the point, intellectually dishonest to believe a claim—for which one admits he is not now able to produce appropriate justification. You will want to say that it is the justification one produces that should convince him of the justifiability of the claim, not his conviction in the justifiability of his claim that should convince him that there must be a justification that can be produced for it.

You will want to say this but you will not. For you have just noticed for the first time that your accuser has in his hands the second edition of Chisholm's *Theory of Knowledge*. Indeed, you notice for the first time that everyone in the waiting room has been reading the second edition of *Theory of Knowledge* and that they all have their books open to pages 16 and 17—the very pages from which I quoted above. And you realize that they have all anticipated what you will want to say and are ready to cite Chisholm in defense of your accuser. For Chisholm has just taught them that there is nothing intellectually dishonest in believing a claim to be true, and believing oneself justified in believing that claim to be true, while admitting, at the same time, both that it is the kind of claim that may well require justification and that one cannot at present produce appropriate justification for believing the claim to be true. On the contrary, Chisholm has just taught them that this is exactly the attitude they should want to adopt toward claims when they conduct a philosophical investigation into the theory of evidence.

Something has gone very wrong here. And what seems to be responsible is a methodological precept that your fellow patients have extracted from the passages in *Theory of Knowledge* I quoted above —a methodological precept they have taken very much to heart. That precept holds that, even if you are very confident that a belief you harbor may stand in need of justification, your inability to produce a satisfactory justification need not undermine your entitlement to persist in that belief. So long as you are confident in the propriety of the belief, you are entitled to believe that, only given enough time and energy, you could elicit an appropriate justification for the belief. That is, you are entitled to believe that your inability to produce such a justification, either now or in the foreseeable future, is of no methodological concern.

It is hardly a precept that we want to be teaching our students. We do not want to be telling them that the critical evaluation of one's cherished opinions is a purely academic exercise. We do not want to be telling them that methodological hygiene requires only that one be confident that one's opinions are justified—that one need not be able to produce the justification, that one need not be able to produce any reason for thinking one's opinions are justified.

But if we are not willing to endorse this methodological precept, then we cannot endorse the way in which Chisholm would reconstrue the question 'What justifies your belief that there is a lectern at the front of the classroom?'—we cannot ask our students to do what Chisholm would have us ask of them. For Chisholm's reconstrual of the question would have us asking them to begin with a belief that they take to be justified and attempt to elicit from themselves their justification for that belief. But then we would be telling them that they are entitled to persist in their belief and to regard their belief as justified even while, engaged in the process of elicitation, they have yet to produce the justification for that belief. And this is precisely what we, fresh from the nightmare scenario, have concluded that we do not want to be telling them. It is because we do not want to have to tell them any such thing that we are unwilling to endorse the methodological precept.

III

To many, I expect that it will seem obvious that Chisholm is being unjustly accused—that his indictment on the charge of promulgating a dangerous methodological precept is the product of some unfortunate confusion. Surely there is a sense of 'justified' in which a person may be justified in believing a proposition even though she cannot produce her justification for believing it. Surely it is just this sense of justification which philosophers have had in mind when they have held that justification is required for knowledge. Surely, were it otherwise, the imposition of this requirement would require drastically revising downward our estimate of the scope of our knowledge. William Alston[6] writes that what plausibility there is to the requirement that a person be able to produce the justification for her justified beliefs is due, at least in part, to "the confusion between 'justification' in the sense of *being* justified and justification in the sense of 'showing that one is justified'. If one fails to keep that distinction in mind," he continues, "one is liable to suppose that in order to *be* justified in believing that *p* one must be *able* at least to 'justify' one's belief that *p* in the sense of showing that one is justified, i.e., exhibiting what it is that justifies one" (*ibid.*, p. 145).

But this is just more dangerous talk. If it is all right for Alston to warn you not to confuse the alleged two senses of 'justified'—if it is all right for him to explain away the plausibility of supposing that a person cannot be justified in believing *p* yet unable to produce her justification for believing *p*—then it is all right for your accuser in the nightmare scenario to do likewise. And it is hard to see how any consideration about how best to analyze '*S* knows that *p*' could possibly bear the weight of our having to say that it is all right for your accuser to dismiss your demand that he be able to produce his justification for believing you a traitor on pain of withholding that belief. The charge of confusion simply provides an apology for the unacceptable methodological precept your fellow patients extracted from Chisholm.

Of course, the mere fact that Alston fails to uncover a confusion that will save Chisholm and his project from indictment, on the charge of promulgating a dangerous methodological precept, does not mean that it is a mistake to think that this indictment is the product of a confusion. Alston's failure shows only that the task of saving Chisholm has to be gone about in some other way.

---

[6] "Level-Confusions in Epistemology," in P. French, T. Uehling, Jr., and H. Wettstein, eds., *Midwest Studies in Philosophy V: Studies in Epistemology* (Minneapolis: Minnesota UP, 1980), pp. 144–5.

What is needed, in order to save Chisholm's project from the indictment, is not a way to defend the precept your fellow patients have extracted from Chisholm's writings, but rather a way to deny Chisholm's commitment to it. What is needed is some way of saying why it is legitimate for our students to persist in their belief that there is a lectern at the front of the classroom even as they acknowledge that they cannot produce their justification for this belief, while it is not legitimate for your accuser in the nightmare scenario to do likewise with respect to his belief that you are a traitor.

It is not too hard to see what we need to say. The precept that one may legitimately persist in a belief and regard it as justified even as one acknowledges that one cannot produce the justification for that belief, we must admit, is indeed a dangerous one as ordinarily understood. It would indeed treat the task of furnishing support for beliefs as an occupation that has no more than academic importance. But, we must contend, this is not how the precept your fellow patients have extracted from Chisholm is meant to be understood. Just as the question 'What, if anything, justifies your belief that there is a lectern at the front of the classroom?' is not meant to challenge the propriety of that belief, so the precept that you may persist in a belief without (yet) being able to produce the justification for it is not meant to entitle you to disregard legitimate challenges to the propriety of your beliefs.

The impressions to the contrary are, we must claim, the product of a confusion. In ordinary contexts, questions about whether one is justified in one's belief are questions about one's intellectual responsibility in holding the belief, about one's entitlement to persist in the belief, about the propriety of one's methodology. But Chisholm is no methodologist. He is not trying to prescribe canons of rational inquiry. He means to fix on beliefs (like the belief that there is a lectern at the front of the classroom) that we already consider to have met all legitimate challenges to their propriety and only *then* to ask what, if anything, justifies those beliefs.

But then, of course, Chisholm can hardly be guilty of promulgating a dangerous methodological precept, he can hardly be guilty of sanctioning your accuser's behavior. Even by his own lights, your accuser's belief that you are a traitor has not met all legitimate challenges to its propriety. Your accuser persists in his belief, indifferent to the fact that his accusation fails to meet what he himself acknowledges to be legitimate challenges to its propriety. And there is nothing in Chisholm's project as just described to give your accuser any comfort in such a posture.[7]

IV

This having been said, we can indeed pose to our students the question Chisholm would have us pose without any fear of instilling in them any dangerous methodological precept, without any fear of disturbing our sleep with the nightmare scenario. There is, instead, a new worry. How are our students supposed to decide upon an answer?

The only source of intuitions concerning justification we have is reflection upon those canons which govern our day-to-day conduct of inquiry and criticism. It is to our practice of inquiry and criticism, in contexts both ordinary and scientific, that we turn in order to arrive at and test our conceptions of good argument and evidential support. But notice that, for the purpose of answering the question Chisholm would have us pose to our students, this source is now completely lost to us. For the canons that govern our day-to-day conduct of inquiry and criticism concern justification as it bears upon proper methodology and intellectual responsibility. And this is just the sense of 'justification' with which, we have had to say, Chisholm is not concerned. As we have had to describe his project, Chisholm would have our students begin their attempt to elicit the justification for a belief only after the belief has already met all the challenges that our canons of inquiry and criticism issue—only after all the guidance offered us by our canons of inquiry and criticism has been exhausted.

---

[7] It should be noted that, while this strategy for extricating Chisholm from the present difficulty works for all of Chisholm's epistemological writings through the 2nd ed. of *Theory of Knowledge*, it does not fare so well for the 3rd ed. There we see a radical departure. The passage from Lewis (see fn. 4), which Chisholm quotes again and again in almost twenty-five years of epistemological writing, is nowhere to be seen. The discussion of how a theory of justification is elicited—the discussion in which Lewis's quotations play a key role for those almost twenty-five years—is nowhere to be found. And suddenly (p. 5) Chisholm is writing that his inquiry into justification is designed, not to acquire self-knowledge, but rather to improve the inquirer's system of beliefs, weed out the false ones, and replace them with true ones.

Chisholm's move to the new position, apparently designed to help him fight off a rising army of externalists (see, e.g., p. 76), is an unfortunate one. After all, what market can there be for the sort of advice Chisholm would now claim to be dispensing? No one needs to be told how to arrive at a justified belief that there is a lectern at the front of the classroom. To maintain that the import of his investigation is methodological is tantamount to maintaining that it has no import at all. And this means that Chisholm's current doctrine would not have offered even an initially plausible answer to my acquaintance's complaint. She wanted to know why it was not a sham to worry for part of a semester over what, if anything, justified her belief that there was a lectern at the front of the classroom. The only answer Chisholm now makes now available, 'Because when we are done you will have important advice about how to determine the presence and location of lecterns', would rightly have made her laugh.

Thus, whatever intuitions we might have hoped to call upon through appeal to reflections upon the conduct of inquiry and the practice of criticism (such as intuitions about the viciousness of certain sorts of regresses or about the defects of circular arguments) will be the wrong intuitions. They will be intuitions about justification in the wrong sense. But what else would we have our students focus upon—what else *is* there to focus upon—to get clear on what justification, in Chisholm's (literally) *extraordinary* sense, is like? How, in the light of the redescription of Chisholm's project we have had to issue, can we expect our students to have the slightest idea about how to answer the question Chisholm would have us ask them?

V

The answer may seem straightforward: by simply telling them more about what sort of answer we are looking for. That, in effect, is just what Chisholm does. To put matters in our terms, Chisholm tells his readers that he is looking for an answer on which the beliefs that are justified (in the extraordinary sense of justification with which he is concerned) are just those beliefs which meet all the methodological challenges we are willing to accept. He tells his readers that he is looking for a regress of justification—unearthing ever earlier links in a (perhaps branching) chain of claims each adduced in support of its successor and terminating in the claim that there is a lectern at the front of the classroom. He suggests to them that the regress can only properly be regarded as a regress of justification if every link in the chain is either immediately justified or has an ancestor link that is immediately justified.[8]

So, if we will but tell our students what Chisholm tells his readers, our students need puzzle over our question no more. Our charge to them will be clear. They are to flesh out a view on which their belief that there is a lectern at the front of the classroom is justified—indeed, on which each belief that meets all the methodological challenges we are willing to accept is justified—by virtue of that belief's being the conclusion of an argument whose ultimate premises are immediately justified for them.

But should we want to tell them what Chisholm tells his readers? His instructions to the uninitiated result, after all, from a number of deliberate choices—none of them incontrovertible nor, indeed, uncontroverted. That we can relieve our students' puzzlement by telling them what Chisholm tells his readers may well provide our students with some comfort. But unless we can see something in the extraordinary sense of justification that underwrites the set of instructions Chisholm chooses to issue, we will be able to afford our students this comfort only at the expense of our own.

---

[8] See, e.g., Chisholm, *Theory of Knowledge*, 2nd ed., chs. 2, 4.

Consider his stipulation that what is justified in the extraordinary sense coincide with what meets all acceptable methodological challenges. It is a stipulation that certainly gives rather short shrift to skeptical concerns. After all, it dismisses from the outset the possibility that there simply is no regress of justification worthy of the name which will support the students' belief that there is a lectern at the front of the classroom. On his stipulation, any account of justification too demanding to admit the existence of a regress supporting the students' belief that there is a lectern at the front of the classroom is *ipso facto* a defective account. For any such account of justification runs afoul of the stipulation's requirement that every claim that meets all the methodological challenges we are willing to accept is to count as justified in the extraordinary sense.

Chisholm anticipates the charge that, by favoring his stipulation over the skeptic's, he begs the question against the skeptic in an arbitrary way. And, disarmingly enough, he pleads guilty. He offers but two lines of defense. The first is that the skeptic is no less guilty of arbitrariness if the skeptic champions some draconian stipulation over Chisholm's. The second is that his (that is to say, Chisholm's) stipulation has at least one virtue that its competitor's lacks—namely, that it accords with common sense. After all, we do think our students are justified in the belief that there is a lectern at the front of the classroom.[9]

To us who are searching for some intuitive grounds for giving our students Chisholm's—as opposed to some other—instructions, the *tu quoque* that constitutes Chisholm's first line of defense can offer little comfort. But Chisholm's second line of defense is likely to make many of us feel more secure. Many, and perhaps most, of us will be inclined to view the choice between defending common sense and attacking common sense to be a fairly straightforward one. We will see the accommodation of common sense to be very much the virtue Chisholm takes it to be.

---

[9] *The Problem of the Criterion* (Milwaukee: Marquette, 1973), pp. 37–8; repr. in *The Foundations of Knowing*, pp. 61–75, see p. 75. Also see *Theory of Knowledge*, 2nd ed., p. 121.

But how can it possibly be a virtue that Chisholm's stipulation possesses? If common sense had anything to say about the extent of justified belief (in the extraordinary sense), there would be no need for Chisholm to issue his stipulation. Recall that the reason Chisholm needs to make his stipulation is precisely because justification in the extraordinary sense is so removed from the realm of common sense that we cannot legitimately expect our students to have any idea of what (in that sense) justifies beliefs. So Chisholm's stipulation cannot have the virtue of according with common sense. His stipulation concerns the extent of justified belief in the extraordinary sense. And that is a matter about which common sense tells us nothing.

Indeed, where common sense does reign—over our everyday canons of inquiry and criticism—there is no issue between Chisholm and his skeptical competitor. All common sense dictates is that there is no methodological challenge worthy of acceptance which impeaches the propriety of the students' belief that there is a lectern at the front of the classroom. All common sense dictates is that there is no regress of justification on pain of producing which the students should withhold belief. But this in no way touches what is at issue between Chisholm and his imagined skeptical rival. Theirs is not a dispute over the scope of justification as it bears on proper methodology. Theirs is a dispute about the scope of justification in the extraordinary sense. And, as we have noted, questions about the scope of justification in this sense have no bearing whatsoever on methodological issues. Thus, the skeptical party in the dispute is as easily able as Chisholm to accommodate the dictate of common sense. He can just as easily as Chisholm grant the authority of common sense on matters within its purview. For on the matter of dispute between him and Chisholm—i.e., on the matter of how many and what sorts of beliefs count as justified in the extraordinary sense—common sense is mute.

Chisholm then has nothing but a *tu quoque* to offer in defense of his stipulation that the class of beliefs that are justified in the extraordinary sense coincide with the class of beliefs that survive all the methodological challenges we are willing to accept: his only defense is that his stipulation is no more arbitrary than any other one might choose.

And this is but the first concession to arbitrariness. The dispute between Chisholm and the skeptic in no way exhausts the extent to which the propriety of Chisholm's instructions to the uninitiated can be, and has been, controverted. After all, Chisholm and his skeptical rival display a large area of agreement. They would both have our students look for the justification of a belief to take the form of a regress of justification, where every claim in the regress either is, or has an ancestor claim that is, immediately justified.

But there are those who would deny that a regress of justification need trace the pedigree of a belief back to beliefs that are immediately justified. Indeed, there are those who would deny that what justifies a belief need take the form of a regress at all.[10] What has Chisholm to say to these critics? What can he say in favor of seeking an account of justification in the extraordinary sense which satisfies his chosen constraints rather than theirs?

One might think that Chisholm's best move is to offer counterexamples to his rivals' proposals, to show how his rivals' proposals betray common-sense intuitions where his own proposal does not. But if such a move really had any force, we would not have found Chisholm in so desperate a position in his dispute with the skeptic. What drives Chisholm to that position is the fact that, once his project is redescribed so as to clear it of the charge of promulgating a dangerous methodological precept, Chisholm can no longer legitimately claim to be taking any position to which common sense commits us. And if Chisholm is not entitled to claim that common sense is on his side as he battles the skeptic over the extent of justified belief in the extraordinary sense, he certainly is not entitled to do so as he fights his nonskeptical rivals over the defining characteristics of justified belief in the extraordinary sense.

Thus, with common sense silenced and the intuitions provided by reflection upon our practice of inquiry and criticism exhausted, Chisholm has no more leverage against these critics than he has against the skeptic. All Chisholm can do in defense of seeking his particular sort of account of justification in the extraordinary sense is point to his rivals and utter a *tu quoque:* all he can say is that the decision to seek his sort of account is no more arbitrary than a decision to seek one of theirs.

But although this sort of defense may succeed in keeping Chisholm's rivals—skeptical and nonskeptical—at bay, it does nothing to alleviate our discomfort. On the contrary, it serves only to make vivid just how high a price needs to be paid to clear Chisholm of the charge that, in reconstruing the question 'What justifies your belief that there is a lectern at the front of the classroom?', he is promulgating a dangerous methodological precept.

---

[10] Adherents to a coherence view of justification would belong in the first group, adherents to a reliabilist view of justification to the second.

For what the concessions to arbitrariness make clear is that the extraordinary, methodologically inert sense of 'justification' with which we have had to say Chisholm is concerned, is a sense of 'justification' with which we can claim no familiarity. Common sense tells us nothing of its nature. Reflection upon the canons of inquiry reveals nothing of its contours. Even the question of whether there *are* any beliefs that are justified in this extraordinary sense is one about whose answer our intuitions offer us no guidance. The concept of justification, in this extraordinary sense, is a creation *ex nihilo.*

But then the theory of justification in this sense, far from being an attempt to answer in a systematic way to our pretheoretic intuitions, is an enterprise that is answerable to none. It is nothing more than an exercise in pure invention and pure stipulation—a game in which every move is legal and each participant plays according to her own fancy. Chisholm's project, redescribed so as to clear it of the charge of promulgating a dangerous methodological precept, becomes instead a project dedicated to uncovering the nature of a concept that has, in fact, no nature to uncover. In having so redescribed it, we have simply traded one nightmare for another.

VI

What has Chisholm done to deserve to be so haunted by the dark realm of restless sleep? What is it about his reformulation of the question that should be held responsible?

One tempting answer is that it is the grip that the foundationalist conception of justification exercises upon Chisholm which deserves the blame. After all, it is because he is in the grip of this picture that he looks for the answer to the question 'What justifies your belief that there is a lectern at the front of the classroom?' to take the form of a regress of justification. And it is because he looks for a regress of justification, because he sensibly does not want to cast aspersion on the propriety of the students' belief that there is a lectern at the front of the classroom, and because he rightly cannot expect them without much labor to be able to produce an appropriate regress that will support that belief, that he must say that they are entitled to persist in their belief and regard their belief as justified even as they seek to elicit from themselves the regress of justification that supports the belief.

And this is just what gets him into trouble. For it is precisely this entitlement—to persist in a belief and regard it as justified even as one admits that one cannot produce the regress of justification that supports the belief—that your accuser in the nightmare scenario claims on his behalf when you demand that he justify his charge that you are a traitor. It is precisely in the granting of this entitlement that Chisholm opens himself to the charge that he is promulgating the dangerous methodological tenet.

This diagnosis of the underlying cause of Chisholm's misfortune is not only plausible. It is also pleasing. The foundationalist picture that so grips Chisholm has, after all, lost its hold on most epistemologists. Other pictures, coherentist and/or naturalist in hue, command much more allegiance from those who write on epistemology today. If it is Chisholm's adherence to foundationalism that is dragging him into black depths of the dreamworld, then perhaps that just shows that there is yet one more reason to do what most epistemologists have already done. To sleep well we need only steer clear of foundationalism and head for one of the epistemologies more in vogue. But which one?

Clearly, it is not to the advocates of coherentism that we should look for relief. Their way of reading the question that caused my acquaintance such distress is in no material way different from Chisholm's. Like Chisholm, the coherentists view themselves as attempting to elicit a regress of justification for our ordinary beliefs about the world. They differ from Chisholm only in their account of the structure a regress of justification is required to exhibit. A coherentist epistemological tract will stand in very well for Chisholm's *Theory of Knowledge* in the nightmare scenario—very well indeed.[11]

In any event, coherentism is not the epistemological doctrine to which most epistemologists appear to have turned. The epistemology of the moment is clearly reliabilism. And the exponents of reliabilism do indeed view their epistemological project in a manner rather different from that in which Chisholm views his. Reliabilists hold that what justifies a belief is its being the result of a process of belief formation that reliably produces true beliefs. It matters not at all whether the person who provides a harbor for the belief believes, or has any reason to believe, that the belief she harbors enjoys this fortunate provenance. Thus, for reliabilists, the search for what justifies the students' belief that there is a lectern at the front of the classroom does not lead, as it does for Chisholm and the coherentists, to a theory of evidence. What justifies their belief is no implicit chain of reasoning—either linear or looped. It is, rather, a fact about the process through which that belief was formed—a fact of which the students might well be unaware.

---

[11] This is nowhere more clear than in Laurence Bonjour's *The Structure of Empirical Knowledge* (Cambridge: Harvard, 1985), pp. 151–2, where he writes: "Thus the basic claim of my coherentist position (and indeed of the analogous claim for most or all nonexternalist versions of foundationalism) must be that careful reflection on actual cases will reveal that these elements [the inferences and belief systems necessary for justified belief] are tacitly or implicitly involved in the actual cognitive state of a person who has empirical knowledge, even though he does not bring them explicitly to mind and indeed would normally be unable to do so even if explicitly challenged" (p. 152).

Consequently, the nightmare scenario that exploits Chisholm's response to the sort of complaint issued by my acquaintance—a nightmare scenario that might equally well have featured the response of a coherentist epistemologist—need raise no goosebumps on a reliabilist's skin. Part of what makes the nightmare scenario nightmarish is that your accuser admits that his accusation is the sort of claim that requires justification, yet he—along with everyone else in the waiting room—is entirely unperturbed by his inability to produce the appropriate justification. The reliabilist, however, refuses to acquiesce in your accuser's admission. For the reliabilist, there is no claim for which a person must have grounds if she is to be justified in believing it. All that is required is that her belief be the result of a reliable true-belief-forming process. (It may happen that the process produces grounds for her belief, but it need not.[12])

VII

But it would not be wise to count just yet on a restful night's sleep. For imagine that it is the reliabilist's epistemological doctrine that is fresh on the minds of your fellow inmates in the doctor's waiting room as you demand from your accuser the justification he has for his outrageous claim. It is true that you need no longer expect that he will admit that his is the sort of claim that requires justification, admit his inability to produce this justification, yet insist that he is entitled to persist in his belief in your treachery and regard that belief as justified just the same.

Rather, you can expect that he will find your demand rather quaint. You can expect him to explain to you that the notion that one needs to be able to provide grounds for one's beliefs in order to have justified beliefs has long been discredited. What makes a belief justified—what makes his belief in your treachery justified—is the fact that it is a product of a reliable true-belief-forming process. His inability to provide grounds either for his claim that you are a traitor or for his claim that his belief in your treachery was produced in the way he described, you can expect him to continue, in no way indicates that he is not justified in believing these claims. Grounds are simply beside the point. You can also expect that, all the while, your fellow patients will be nodding in agreement.

---

[12] See, e.g., the reliabilist theory of justification defended by Alvin I. Goldman in "What is Justified Belief?" in *Justification and Knowledge*, George S. Pappas, ed., (Dordrecht: Reidel, 1979), pp. 1–23.

The nightmare is all too familiar. Once again, your fellow patients have extracted from a reputable epistemological doctrine a methodological precept that would sanction your accuser's reply to your demand that he produce grounds for his outlandish accusation. Once again, it is a dangerous methodological precept. And, once again, it looks as if we cannot ask our students to endorse the reputable epistemological doctrine from which the dangerous methodological precept has been extracted without thereby promulgating that precept.

But this time it is reliabilism, not foundationalism, that is to blame. It is from reliabilism that your fellow patients have extracted the precept that, since it is the reliability of the process producing a belief which renders the belief justified, and since one's being able to meet legitimate methodological challenges to a belief is neither necessary nor sufficient for its being produced by a reliable process, one is free to dismiss any such challenges; to the question of whether one's belief is justified, one's ability to meet such challenges is irrelevant. This methodological precept is no less dangerous than the one extracted from Chisholm. And it looks to be no less implicit in the reliabilist account of justification than the previous precept was in Chisholm's.

Not that there are no responses available to the reliabilist. Like Chisholm, he can claim that his doctrine is not meant to have methodological import,[13] and thus that his answer to the question of what justifies a belief can provide no comfort to your accuser. He can even claim, if he wishes, that the question he means to answer arises only after all legitimate challenges to the belief have been satisfactorily met (or, that the question arises only for beliefs against which no legitimate methodological challenges can be raised). But these responses can get him no farther than they got Chisholm. Once the reliabilist denies that his use of 'justified' is to be understood in the ordinary way, he relinquishes his right to call upon ordinary intuitions concerning the nature and extent of justified belief in support of his theory of justification. And without recourse to such intuitions, his theory can be nothing more than an exercise in pure stipulation.

---

[13] As Goldman does in *Epistemology and Cognition* (Cambridge: Harvard, 1986), pp. 25–6, and in "The Internalist Conception of Justification," in *Midwest Studies in Philosophy V: Studies in Epistemology*, p. 28.

VIII

There are some writers, reliabilist and nonreliabilist, who will undoubtedly protest this description of their predicament. They will claim, as we have already seen Alston do, that the distinction between uses of 'justification' to which they appeal and which I have so deprecated is, in fact, underwritten by our ordinary intuitions. But this claim, of itself, cannot reverse their fortunes.

What anchors our intuitions about the nature of justification as we ordinarily understand it is the role ascriptions of justification play in the conduct of inquiry and criticism. Questions about justification, as the term is ordinarily understood, are questions about the propriety of methodological commitments and the propriety of adopting (retaining) the beliefs these commitments sanction. These are consequential questions. To conclude that a belief has not the propriety we supposed is to demand that it be denied the credence it has hitherto been accorded. It is the fact that we have an actual practice of inquiry and criticism at stake when such questions are raised which places constraints on what answers we can, in all honesty, take seriously.

To uphold the claim that ordinary intuitions support the existence of some other sense of justification, something comparable needs to be said about what anchors intuitions about justification in this other sense. Something needs to be said by way of describing the endeavor in which ascriptions of justification in this other sense are supposed to play a role—something to convince us that our intuitions do indeed place substantive constraints on theories of justification in this other sense. But what is there to say?

Some have said that it is our intuitions about knowledge that anchor our intuitions about the concept of justification in which they are interested. It is justification, in the sense that justification is necessary for knowledge, into whose nature they mean to inquire.[14] But this is an appeal we can safely set aside. If Gettier has taught us anything, it is that, after we have determined that *P* is true and decided that an investigator has met all legitimate methodological challenges to her belief that *P* (i.e., all challenges she must meet on pain of having to withhold belief that *P*), the question of whether her belief counts as knowledge is still open. One lesson, then, is that, whatever is at stake in how we answer that open question, it has nothing to do with the propriety of the manner in which the investigator has conducted her inquiry; its propriety has already been settled. But that is just to say that we are free to answer the question of whether her belief counts as knowledge however we like—we are free to adopt any position concerning the nature and extent of propositional knowledge—without fear that we will have to pay a price for our stand, that there will be something in our conduct of inquiry or anything else that will have to be trimmed to fit. And that means that calling upon intuitions about knowledge to anchor intuitions about justification cannot deliver the help needed. For it means that intuitions about knowledge are themselves adrift from any and all human endeavors that might conceivably constrain them.[15]

---

[14] For such an appeal, see Alston, "Level Confusions in Epistemology," p. 144; Goldman, "The Internalist Conception of Justification," p. 29.

The only other gambit of any promise seeks to exploit an apparent analogy between the epistemic and moral realms. When contemplating an agent who must make a moral decision, we can ask 'How ought that agent to act?' There are at least two different things we can mean. We can mean to be asking: 'What does the correct moral theory require the agent to do in her circumstances?' Or we can mean to be asking: 'What does the correct moral theory require the agent to do given what the agent takes to be her circumstances?'[16] Meant in the first way, the question asks which among the available courses of action is *objectively right*—if act utilitarianism is correct, which of the available courses of action will maximize happiness. Meant in the second way, the question asks which among the available courses of action is *subjectively right*—if act utilitarianism is correct, which of the available courses of action would maximize happiness were the world the way the agent takes it to be.

Objective and subjective rightness often fail to be satisfied by the same acts. An act-utilitarian agent who has misleading information about her circumstances may well perform an act thinking it will maximize happiness when, in fact, it will not. If act utilitarianism is the correct moral theory, her act will then have been subjectively right but objectively wrong. It is clear from this case that the latter assessment is far removed from regulative concerns. The verdict that her act was objectively wrong attaches no taint to the agent—it is compatible with the verdict that no one in her circumstances could have done any better. It is in our assessment of the subjective rightness of the agent's act that we express regulative concern—our concern about whether she did the best she could in her circumstances, whether she ought to take steps to improve her performance. But the two assessments, subjective and objective, are this much connected: an agent's act is objectively right if and only it is an act she would be prepared to perform were she to know and act upon all the moral and nonmoral facts relevant to her decision.

---

[15] I offer a fuller discussion of these matters in "It's Not What You Know that Counts," this JOURNAL, LXXXII, 7 (July 1985): 350–63.

[16] A third thing we can mean, but which we can ignore for the present discussion, is: "What is the agent required to do given what she takes to be the correct moral theory and what she takes to be her circumstances?" For discussion of this way of taking the question, see Hilary Kornblith, "Naturalizing Rationality," in Newton Garver and Peter H. Hare, eds., *Naturalism and Rationality* (Buffalo: Prometheus, 1986), pp. 115–33.

Our intuitions, it has been claimed, support an analogous distinction in epistemology. Faced with an agent harboring (or about to harbor, or about to abandon) the belief that *P*, we can ask 'Is the agent's belief justified?' Again, it is claimed, we can mean at least two things. We can mean to be asking whether the agent's belief is *objectively* justified or whether it is *subjectively* justified. And again, the story goes, it is only meant the second way that the question expresses a regulative concern—a concern about whether the agent has met all the challenges to her belief that the proper methodology demands. When meant the first way, the question is nonregulative, devoid of methodological import. And, it is claimed, it is to our intuitions about justification in this first, objective, sense that the reliabilist theory of justification means to answer.[17]

But the claim cannot be sustained. Suppose we let the analogy between the moral and epistemic realms guide our intuitions. We should then expect, not only that the two ways of assessing justification should express different concerns, but also that they should exhibit the same sort of connection we see in the moral case. We should expect that an agent's belief will be objectively justified if and only if it is a belief she would hold were she to know and act upon all the epistemic and nonepistemic facts relevant to the adoption (retention) of the belief. But, given that truth is a central aim of inquiry, one key fact relevant to the adoption (retention) of an agent's belief must be whether the belief is true. So, following the analogy, we should expect that an agent's belief will be objectively justified if and only if it is a belief she would hold if she knew the truth value of the propositional object of the belief. And from this it follows that all and only true beliefs are objectively justified. But this is a condition that no reliabilist (or foundationalist or coherentist) appears even slightly inclined to take seriously.[18]

---

[17] This appears to be like Goldman's view. See his *Epistemology and Cognition*, p. 25 (see fn. 4); and "The Internalist Conception of Justification," pp. 28–9.

One could, of course, contrive to loosen the relation between the criteria of subjective and objective moral rightness, or to loosen the analogy between the moral and epistemic realms, or both. But, it is precisely the tightness of the relation of objective to subjective moral rightness—the fact that the assessments of objective moral rightness are just assessments of subjective moral rightness made in the light of knowledge that may be unavailable to the agent—that lends the concept of objective moral rightness *its* intelligibility, *its* anchor in matters of practical import. And it is the tightness of the analogy between the moral and epistemic distinctions which is supposed to lend the concept of objective justification its intuitiveness. To engineer this sort of escape from the consequences of the analogy drawn is to forego any intuitive cachet the analogy might have afforded the enterprise of constructing a theory of justification, where 'justification' is understood in any sense out of the ordinary.

Thus, if objectively justified belief is not just true belief, then the appeals to intuitions about knowledge and about objective justification offer no relief. Unless something has been overlooked here, there is simply no reason to think that our intuitions give us purchase on any extraordinary notion of justification at all.

IX

The tempting, happy diagnosis of the cause of Chisholm's misfortune is, thus, mistaken. It is not Chisholm's adherence to foundationalism that explains his encounter with the fearful world of dreams. Chisholm's coherentist and reliabilist counterparts have to contend with much the same nightmares. Rather, all the nightmares have their origins in one and the same assumption—an assumption that Chisholm shares with his coherentist, reliabilist, and even skeptical colleagues. The assumption is that, even after our students have satisfied us that they have done everything that intellectual responsibility and methodological hygiene demands of someone who believes that *P* (in their circumstances)—that they have met all the challenges that need to be met on pain of withholding the belief that *P*—we can still sensibly say to them, "Yes, yes, but you haven't yet finished saying what (if anything) justifies your belief that *P*." The assumption is that there is something more to say.

---

[18] John Pollock offers much the same argument (albeit in service of a somewhat different cause) in "Epistemic Norms," *Synthese*, LXXI, 1 (April 1987): 61–95; see pp. 77–8.

But when we have examined what more there is supposed to be left to say, it has invariably turned out to be something that, read flatly, offends against uncontroversial tenets of sound methodology and, read with the aid of the sort of distinction required to prevent the offense, succeeds in so doing only by employing some sense of justification about which we can claim no intuitions at all. We could continue our survey of what more there is supposed to be left to say, in the hope of finding some account that will succeed where foundationalism, coherentism, reliabilism, and skepticism have failed. But there would be no point.

We have already learned that, on pain of losing our intuitive bearings entirely, we have no choice but to read our talk about justification flatly. But, read flatly, we cannot claim that our students have not yet finished saying what (if anything) justifies their belief that *P* without thereby accusing them of having failed to have done all that methodological hygiene and intellectual honesty demand of someone who believes that *P*. Thus, once we have conceded that our students have met all legitimate methodological challenges to the propriety of their belief that *P*, we have conceded that they are finished saying what justifies their belief that *P*.

Small wonder then, that, in attempting to tell our students how to say something more, Chisholm and his coherentist, reliabilist, and skeptical competitors should end up saying things that, read flatly, run afoul of methodological precepts. Small wonder that there should be no hope that we might shop around and somehow find a better way to instruct our students on how to say something more. There is nothing more to say. There is nothing more to say in response to the question 'What justifies your belief that *P*?' than intellectual honesty and methodological hygiene demand that we be able to say on pain of withholding the belief that *P*.

And, if so, my acquaintance's complaint turns out to have been indeed well-taken. She found her instructor's challenge to say what, if anything, justified her belief that there was a lectern at the front of the classroom to be a sham, a display of intellectual dishonesty. Those are harsh accusations, but she turns out to have been this far right: the question admits of no interpretation under which it deserves, or can receive, a serious answer.

Those familiar with Chisholm's corpus will recognize the dialectical significance of this result. The passages I quoted at the outset were not, after all, written in response to my acquaintance's complaint. Chisholm wrote them in response to philosophers of pragmatic and/or Austinian and/or Wittgensteinian bent who had argued that asking someone for the justification of an everyday perceptual belief is perverse and/or an abuse of language.[19] Chisholm's strategy, recall, was to admit the force of their objections to the question as ordinarily understood but to suggest that there is room to understand the question in a different way. Chisholm's ambition was thus to undercut the pragmatic/Austinian/Wittgensteinian critique to which my acquaintance simply gave a naive voice.

What we have seen is that Chisholm's ambition cannot be realized, his strategy cannot succeed: there is no room to understand the question Chisholm wants to ask in anything but the ordinary way. The perversity that my acquaintance saw in the question her instructor asked by way of drawing his class into the concerns of contemporary epistemology—the perversity to which Peirce, Austin, and Wittgenstein tried to draw attention—cannot be finessed away. The question that Chisholm wants to ask—the question that his foundationalist, and his rivals' coherentist, reliabilist, and even skeptical theories of justification would answer—is fundamentally misguided.

Were it just the question that deserved the blame, this neo-pragmatic/Austinian/Wittgensteinian storm might yet be weathered. After all, the mere fact that the theories of justification we have canvassed can be regarded as attempts to answer a misguided question—even the fact that at least some of these theories have been regarded from the outset as attempts to answer this question—hardly means that they must be so regarded. Were it just this question that deserved repudiation, we might still be able to view these theories in some other, more favorable, light—as the products of some other, better motivated enterprise. And if so, foundationalism, coherentism, reliabilism, and (perhaps) even skepticism might survive the foregoing critique and retain their claim to be competing theories of genuine interest.

But if there really were something about which foundationalism, coherentism, reliabilism, and skepticism could properly be said to be competing theories, there would be no reason to cast aside Chisholm's question. We could simply interpret Chisholm's question as that question to which these theories offer competing answers. The reason we have deemed Chisholm's question misguided is that we have learned that, unless we want to say that these theories are promulgating dangerous methodological precepts, we cannot but conclude that there simply is no pretheoretic concept of justification about whose nature they can properly be said to differ.

---

[19] See the passages cited in fn. 1 and 2.

It is not Chisholm's question, but rather the vision of the theory of justification which he shares with his coherentist, reliabilist, and skeptical confreres which is the problem. Theirs is a vision of a pure theory of justification, separated from all that would make it methodologically potent, cleansed of the concerns proper to the realm of the ordinary. The problem is that a theory of justification thus purified, a theory of justification deprived of any role in methodology or the conduct of inquiry and criticism, is a theory that divorces epistemology from the very practices that furnish it with its only source of intuitive constraint. It is epistemology on holiday.

We need not fear that, in recognizing this problem and rejecting the vision of epistemology that is responsible, we will have turned our back on the question of what justifies belief. If there is nothing more to say in response to the question 'What justifies your belief that *P*?' than methodological hygiene and intellectual responsibility demand we be able to say on pain of withholding the belief, then the question that deserves our attention is 'What exactly do methodological hygiene and intellectual responsibility demand of us?'

There is in philosophy a noble tradition of theorizing about the nature of justification which addresses just that question—which seeks to clarify, to criticize, to improve the conduct of inquiry and criticism. It is a tradition exemplified in the seventeenth century by Descartes (*Rules for the Direction of the Mind*), in the first half of our own century by the authors of logical empiricism, and in contemporary writing (I would argue) by advocates of the Bayesian approach to epistemology and the philosophy of science. And this is to mention but three instances in which philosophers have undertaken to confront deep methodological issues and evaluate the way in which inquiry is properly to be conducted. Much more of the very same sort of work beckons us today. Which is to say that the theory of justification can survive very well the loss of the idle question in whose name it has journeyed so far and to such little effect. Epistemology need not stay on holiday; it has a job waiting at home.

MARK KAPLAN

University of Wisconsin/Milwaukee

BOOK REVIEWS

The Right to Private Property. JEREMY WALDRON. New York: Oxford University Press, 1988. viii, 470 p. Cloth $59.00.

The Right to Private Property displays intellectual and literary virtues not often found together—boldness and sobriety, sharp focus, and broad sweep. Aware of the diffuseness of the topic of property rights, Jeremy Waldron has chosen to focus narrowly enough to ensure intellectual discipline. He concentrates only on "rights-based" arguments for private property, leaving to one side and another occasion utilitarian arguments, arguments from the preservation of distinctively "republican" political virtues, and "Aristotelian" natural-law arguments from the fostering of desirable sorts of character. By the same token, he sticks to arguments over private property, and even where these lead, as they do with Proudhon, to the conclusion that "property is theft," he does not speculate about what forms of common or public ownership a rights-based approach would license.

Yet, the first hundred and twenty pages of *The Right to Private Property* are devoted to the two largest issues in discussions of property rights, namely, the nature of property rights and the nature of rights-based arguments; and even though the discussion of rights-based theories concentrates on a handful of authors—Locke, Robert Nozick, Kant, Hegel, and Proudhon—the conclusions toward which the argument drives are of the broadest. Reversing the trick of many authors, who offer as a conclusion what is essentially an introduction, Waldron spells out his own conclusions in the first couple of pages.

Starting from the rhetorical claim that property is a human right, he asks "whether there are any good right-based arguments for private property." A right-based argument is understood in much the way Ronald Dworkin understands it, that is, as an argument from an *individual* interest to the conclusion that the interest is of such importance that government is bound to promote it. A right-based argument for private property is thus an argument to the effect that governments are obliged to maintain a system of private property as opposed to communism or some other alternative. Such arguments come in two forms. The Hegelian argues that owning property contributes immensely to the ethical development of the individual, the Lockean that those who have acquired the appropriate interest in an object, by labor or legitimate transfer, have an interest that governments are obliged to protect.

0022-362X/91/8803/155–9

The Hegelian argument thus presupposes a general right, the Lockean a special right. Waldron's argument is that the Lockean theory of acquisition, like the refurbished version offered by Nozick, is unconvincing, because there is no plausible justification of the transition from mixing one's labor to acquiring what amounts to "full liberal ownership"; nor is there any plausible justification of the obvious alternative, the argument from "first occupancy." Locke emerges with rather more credit than Nozick, because he assumes a background of a general right to subsistence which constrains the acquisition of would-be owners of property, and Waldron argues with considerable plausibility that "no theory of the kind that Nozick intimates could possibly be made acceptable in the absence of a back-ground general right to subsistence."

In the case of Hegel, the issues are different. Hegel demands that all adult persons be property owners; the argument—that only thus can they show the responsibility and autonomy that define moral agency as we conceive it—is tortuous and difficult, but engaging. Waldron does not want to claim more success for the Hegelian enterprise than he can demonstrate, but agrees that it establishes some connection between personality and ownership; perhaps more to the point, the Hegelian tradition has done it by paying attention to social and political phenomena that appear in libertarian writings only at the level of rhetoric.

Once we accept even so much of a Hegelian perspective, interesting and difficult issues abound. The most obvious is the question whether the regulation and redistribution required to ensure that *everybody owns property* will not so cut into the rights of owners to do as they choose with their own that nothing resembling private property is left. Waldron's view is that most people who claim that this would be the result of taking Hegel seriously just go too fast. Certainly, it might be true but, on the face of it, it need not be. There are numerous different conceptions of private property, and by no means all of them are inconsistent with a good deal of regulation and redistribution. Nonetheless, the egalitarianism implied by taking Hegel seriously cannot be absolute. Acknowledging the merits of private ownership is inconsistent with the suppression of all inequality. More contentious is the conclusion that acknowledging the right to property is equally inconsistent with accepting the degree of inequality manifested in capitalist economies where a few people have lots of property and lots of people have none whatever.

As this summary suggests, Waldron is unafraid of contentious conclusions. What it cannot suggest is the continuous stream of interesting and novel argumentation on the way to those conclusions. Waldron's discussion of the definition of private property, for instance, is a splendid example of brisk common sense allied to considerable subtlety of analysis. It helps that he is a lawyer as well as a philosopher, and can bring a great variety of legal material to bear on our first intuitions about the essence of private ownership. Acknowledging that many writers have despaired of defining property, for the not wholly disgraceful reason that the objects of ownership, the nature of their owners, and the legal relations between owners and what is owned have varied very widely, Waldron insists that his subject cannot properly be discussed without some pretty clear account of just what private property amounts to. Following H.L.A. Hart and John Rawls, he suggests that variations in the various dimensions of ownership can be handled by thinking of different conceptions of the one concept of private property.

As to the core notion of property, he starts from the thought that property rules are rules governing access to and control of material resources. The now famous instance of why the rule forbidding me to stick my knife in your chest seems not to be a rule about property, even though it governs my use of a material object, is easily handled; my *ownership* of the knife is a matter of the distribution of access to and control over the knife, but the rule forbidding me to stick it, or any other knife, in your chest is not. The difficulty presented by the English Common Law's doctrine of estates, which held that what the owners of land really owned was an incorporeal estate *in* the land rather than the material stuff itself, is set aside; strictly speaking, there was no private ownership of land in English feudal law. Talk of incorporeal estates is just a way of talking about a form of collective ownership.

Applied generally to the phenomena of the ownership of shares, copyrights, mortgages, and reversions, Waldron's approach is to stick firmly to the idea that property rules govern access to and control over material resources, and that where, as in a modern economy, we have widespread corporate ownership of the material resources themselves and private ownership of shares in the corporation, we should say that the man who owns *shares* does not own material resources, even though his wealth does indeed lie in the rules that determine access to those material resources. This is thoroughly sensible, but leaves me at least unsure just what the upshot is. Is 'ownership' equivocal as between owning things and owning claims on things, are such shares as I may own *not* my property, or what? The problems that lead so many of us to say that ownership is whatever counts as ownership in a given legal system have not entirely been exorcised. If the definition of property causes trouble, the privateness of private property does not. A property regime is a private-property regime when it puts all the incidents of ownership into the hands of nameable individuals. The contentious conceptual issues have to do with how many of these incidents and which of them it takes to make it a system of private ownership, on whichever of those two words the stress is taken to fall.

The discussion of particular writers which follows these preliminaries is continuously interesting. Devotees of Locke might complain that Locke is too closely assimilated to Nozick, and admirers of Hegel might wince occasionally when Waldron presses on a favorite equivocation in the master's exposition of the individual's need to manifest his will in ownership, but overall the discussion is a model of fair-minded exposition and criticism. All the same, this is well-tilled soil, as Waldron generously acknowledges, both in the text and in his footnotes. Although Waldron's treatment of the subject is by some way the best now in print, he has had a bit of help from those who came before. Let me conclude, therefore, by mentioning one writer whose work has never been treated with the respect that Waldron accords it, and where Waldron turns all one's expectations upside down.

In twenty pages, Waldron rescues Proudhon from a century and a half of obloquy. Most twentieth-century readers who know anything at all of Proudhon know him only as the author of the famous observation that "property is theft," and thus as the target of the obvious putdown that theft presupposes property since nobody can steal what nobody owns. In Waldron's account, Proudhon simply held that we all have an equal general right to become an owner of property. This supports a positive view of the morality of proprietorship in some shape, but also gives sense to the claim that anyone whose holdings make it impossible for another to become a property owner has deprived him or her of their property, and is therefore guilty of theft.

Kant had earlier come close to such a claim when attacking feudal obstructions to the easy alienability of landed property, but Proudhon was more impressed by the thought that, if we take seriously the right of everyone to equal liberty, *and* the fact that that liberty demanded property for its expression, we must curtail the rights of existing owners for the sake of those otherwise left propertyless. What is appealing about Waldron's treatment of the argument is the deftness with which he distances himself from Proudhon's own formulation, which does indeed suggest that in curtailing the usual rights of owners we are *abolishing* property, and then uses Proudhon's insight as the basis for an egalitarian account of the distributive implications of a general right to property. It is not surprising that Waldron draws one of the epigraphs to *The Right to Private Property* from Proudhon. "Whenever a person is born, the others must crowd closer together"—the haves cannot simply insist on their rights of ownership against the right to liberty and the welfare needs of the have-nots.

ALAN RYAN

Princeton University

Fact and Method. RICHARD W. MILLER. Princeton: Princeton University Press, 1988. 611 p. Cloth $65.00, paper $14.50.

Philosophy of science often appears as a subdivided estate, its products the work of specialists. Publications, whether articles or books, tend to focus on one or another of the subtopics such as explanation, confirmation, or realism. In this context, it is good to see a new book such as Richard Miller's *Fact and Method,* in which all the major issues are treated with the care and coherence to show their close interdependence. This is by no means a piecemeal survey of the philosophy of science; rather, it is a forceful and cogent review of scientific methodology which weaves a tough thread of relevance through the analysis of explanation, confirmation, and realism.

Fact and Method, with its ambitious goal of replacing positivism as the dominant perspective on science, provides a clear and accurate atlas of the discipline of philosophy of science. As the author argues in detail against the positivist account and in favor of his replacement, we see clearly why the analyses of explanation and confirmation are essential groundwork for an argument for realism, the author's dominant cause. Accurate descriptions of these scientific activities function as the evidence for claims about the appropriate level of justification. Thus, with the author's demonstration that the complete argument for realism must be "book-length," he acknowledges the need to understand what science does before we can adjudicate how much of it is believable. And the effect of this good strategy is a philosophy-of-science book with a breadth of perspective to make it readable and informative both for the general philosophical audience and for philosophers of science.

The case against positivism begins with an account of the failure of the covering-law model of explanation. This is the core of the posi-

0022-362X/91/8803/159–62

tivist's approach to explanation and it fails to describe adequately that scientific activity because it does not attend to the causal aspects of the phenomena to be explained. A covering-law explanation shows why an event was to be expected but not necessarily what brought it about. Capturing the important features of scientific explanation, Miller argues, requires analysis as, "an adequate description of underlying causes" (60). This demands a clear concept of what it is to be a cause and an indication of how we are to recognize causes. Miller responds to this requirement by treating causality as a "core concept." Our understanding of a core concept (others being number and work of art) is based on a diverse pool of exemplars and procedures for extending the concept so that other examples can be judged as being relevantly like the exemplars. There is no general analysis of the concept. Specifications of causality will be discipline-specific in the sense that the sciences themselves must specify the exemplary causes and extension procedures. If theories of interactions change within science, so too will the concept of cause. Thus, the adequacy of a causal account, that is, the nature of explanation, is a scientific, not philosophical decision.

This reveals the principle methodological tack in *Fact and Method*, a counsel against general, a priori (philosophical) accounts of science and an emphasis instead on case-specific (scientific) principles. The pervasive flaw in positivism is its reverence for generality and its insistence on rules of scientific reasoning which span the sciences. This is the germ of its inadequacy. Miller's demonstration against this kind of generality is supported by a wide variety of examples wherein each scientific discipline fills in its own causal rules and thus sets its own explanatory guidelines. The examples in this case and throughout the book show a rich diversity of sciences, both social and natural. This dimension of breadth is another strength of the book.

The causal account helps us understand a necessary feature of explanation. But more epistemological work needs to be done in weeding out unjustified assessments of cause. This is the task of confirmation. Miller's diagnosis of positivist accounts of confirmation (hypothetico-deductive and Bayesian models alike) finds the same deficiency in their blindness to causal factors. The remedy is to cast confirmation of a hypothesis as, "fair causal comparison with its rivals" (155). That is, a hypothesis is confirmed by explicit comparison with actual alternatives, arguing from principles shared by all sides. One accepts the hypothesis that is entailed by the best causal account of the data and our beliefs about them. The evaluation covers not just the account of how the phenomena in question arise, but also the story of the interactions by which we observe the phenomena and how we come to entertain the belief. Miller provides mainstream examples (episodes in confirmation of Darwin's theory and Newton's) to illustrate how this model of confirmation maps accurately onto real science.

This coherent treatment of explanation and confirmation lays the important foundation for Miller's argument for realism. Since acceptance of a hypothesis is linked by the account of confirmation to the causal properties of the entities involved, it is also linked to truth. That is, the virtues that mark a hypothesis as the best explanation are truth-conducive virtues and not simply psychological (the phenomenon was to be expected) or pragmatic (the hypothesis is the simplest). A causal explanation has the existential claim built in, in the sense that, if you confirm the explanation involving entity *x*, then you confirm as well that *x* exists, since "what does not exist cannot cause" (382).[1]

This neat link between best explanation and truth seems too good to be true, and it is. If the account of explanation and confirmation is to carry into the fray of realism sufficient momentum to change skeptical minds, it will require significantly greater epistemological precision in describing causality. The fuzzy notion of cause as a core concept will suffice in the argument that explanation requires a causal account. But this half-analyzed concept does not provide the necessary link between our ability to recognize a cause, based on a dynamic, topic-specific core of exemplars and extension procedures, and truth in causality. The specific realist case that entity *x* exists relies not only on the truth of the conditional that, if *x* is cause, then *x* exists, but also on our ability to establish the truth of the antecedent. It is unanswered doubt about this ability which at this point leaves antirealism as an option.

But Miller's case for realism is a much richer and more detailed argument than I have indicated so far. He is specifically opposing an antirealism of that dodgy form which maintains only that an attitude of agnosticism toward the existence of unobservables is always a reasonable option. The rebuttal to this comes through specific cases (as it must, to be in line with the methodological opposition to generality) in which "topic-specific truisms" leave the scientist no reasonable alternative but to believe the particular explanatory theory. Examples of such principles that "can move all reasonable people" (469) are given for the theory of microbes and for atomic theory as an explanation of Brownian motion. In the latter case, it is simply unreasonable to deny that the acceleration of the particles is caused by external forces. Failure to endorse such a truism is failure to participate in physics. Believing such a truism is in this sense a criterion of rational participation in science.

---

[1] This is much like Nancy Cartwright's analysis in *How the Laws of Physics Lie* (New York: Oxford, 1983).

In this way, philosophy learns from science itself the standards for what is and what is not reasonable to believe. Scientists, that is, "have something to teach philosophers" (471) about realism. It will, of course, be a convincing lesson only for those who are predisposed to respect the scientist's judgment of believability. Miller acknowledges (384) this presumption of respect for science but does not consider it a serious weakness in the case for realism. But the two prerequisites, this respect of science and a definitive epistemology of causality, puts *Fact and Method* in place with most other voices in the cause of realism, compelling only to realists. Miller's contribution, though, is among those most worth hearing since its book-length argument, by weaving explanation and confirmation into a complete case, nicely demonstrates how the argument ought to go. But book-length or not, this argument is unlikely to convert the nonbeliever since its points of fuzziness will always turn out to be in the critical view.

And there is irony in Miller's methodological persuasion, which he shares with many others, to direct philosophy of science toward the specifics of science and away from the generalities of philosophy. As logical positivists marshalled metaphysical arguments to ban metaphysics from philosophy of science, Miller presents a general account that is intended to pre-empt generalizing. The causal account of explanation and confirmation avoid the positivist's generality, he claims, because it "describes jobs that must be done, but not the content of the principles that do them" (199). But clearly the description is meant to apply generally in the sense that all explanation is to speak of causes. Much like the covering-law model, Miller's account prescribes a general structure that leaves room for specific content. The real difference is that the positivists tend not to fill the general model with specific principles. They think that general claims complete the understanding of science. For Miller, on the other hand, we do not fully grasp what it is to be an explanation (for example) until we fill the general form with specific content. The persuasion then should not be so much against generality as *for* specifics. Generalizing, after all, is the accomplishment of understanding. Just as science understands the world through unification and generalization, philosophy understands science with similar style. So it is true: scientists do have something to teach philosophers.

PETER KOSSO

Northern Arizona University

Explaining Science: A Cognitive Approach. RONALD N. GIERE. Chicago: University of Chicago Press, 1988. xxi, 321 p. Cloth $34.95, paper $17.95.*

Most recent discussions of problems in the philosophy of science take a highly idealized view of scientists. Insofar as cognitive agents can be discerned at all behind the staples of philosophical analysis—the reconstructed theories, the confirmation functions, the covering-law explanations, and so forth—they are logically omniscient systems with infinite memory, who make instantaneous responses to stimuli and who act in isolation from one another. For many other students of the scientific enterprise, particularly historians and sociologists of science, the philosophical depiction of the scientist is so idealized as to lose any contact with the phenomena. Thus, in at least some quarters, philosophy of science is dismissed as a scholastic irrelevancy, a cluster of intricate games pursued with odd obsession by some apparently highly intelligent people.

Ronald Giere hopes to remedy this unfortunate situation by adopting a "cognitive approach" to science. He proposes to use ideas from contemporary cognitive science to reformulate some philosophical problems about science and to reconnect philosophy of science with the research of historians and sociologists. Left to their own devices, historians and sociologists pursue philosophy by other means, and, Giere believes, some of their philosophizing, particularly the flirtations with relativism, needs critical scrutiny. The cognitive approach is not only to correct the scholastic excesses of philosophy but to trim the ambitions of sociology as well.

Giere begins with two chapters that present his ideas about the current state of science studies and of epistemology generally. He offers thumbnail sketches of philosophical naturalism, of evolutionary epistemology, of logical empiricism, and of the work of Thomas Kuhn, Bas van Fraassen, and such contemporary sociologists as Barry Barnes, Harry Collins, and Bruno Latour. This is followed by chapters in which he formulates his own version of the semantic conception of scientific theories, using it to specify a position that he calls "constructive realism." The central idea is that scientists typically (and correctly) contend that the world is similar to specified models in specified ways and to specified degrees. The realist emphasis comes in with the contention that the ways in which models are

---

* I am extremely grateful to Gary Hardcastle for some valuable discussions and his constructive comments on an earlier draft.

0022-362X/91/8803/163–7

akin to reality are not confined to agreement on observables: typically, scientists will claim approximate agreement in a far wider variety of respects.

In his fifth chapter, Giere attempts to defend constructive realism by appealing to his experiences at the Indiana University Cyclotron Facility (IUCF). His targets include empiricists (primarily van Fraassen) and social constructivists (such as Karin Knorr-Cetina), whose views he opposes by reporting how firmly his physicist interlocutors seem to be committed to the reality of the protons that they accelerate.

In chapter 6, Giere's focus changes. He outlines a decision model for scientists, drawing on well-known psychological studies that question the view that scientists are Bayesians. Giere's alternative is to claim that scientists are satisficers, and he offers an account of how the practice of testing can be interpreted if we view scientists as facing a satisficing decision problem. The two final chapters are intended to explore this idea further. Giere uses his own "ethnographic" findings, based on interviews with research scientists, to analyze work in nuclear physics and in plate tectonics from his own perspective on scientific decision making.

Does all this amount to a cognitive approach to the philosophy of science? Hardly a fully-articulated theory—but perhaps, as I shall suggest below, a program for future development. For much of the book, the appeals to the literature of cognitive science are quite disconnected from the philosophical discussions. Periodically, readers are offered expositions of various experimental results whose relevance to the philosophical problems is never made clear. Most importantly, Giere has no overall picture of the cognitive subject sufficiently detailed to inform his ideas about theoretical knowledge and about decision making.[1] As we shall see, the absence of any such account leads to crucial vagueness in Giere's own treatment of scientific decision.

In my judgment, the most successful part of the book contains Giere's exposition of the semantic conception of theories. Here he uses an illuminating example from textbook physics (the linear harmonic oscillator) to reveal how a small part of classical mechanics can be represented from his favored perspective on theories. Moreover, he is able to show how the conception can do useful work in distinguishing various kinds of realist or empiricist commitments. At this point, however, the picture blurs. Giere's own efforts to advocate constructive realism seem to me to be flawed in important ways.

---

[1] There is a useful contrast here with philosophers who do employ a general perspective on human cognition to reformulate philosophical questions. See, for example, Alvin Goldman's *Epistemology and Cognition* (Cambridge: Harvard, 1986).

According to Giere, use of the semantic conception of theories frees the realist from all kinds of traditional problems about truth and reference. Instead of worrying about correspondence between linguistic (or mental) representations and the world, we can attend to the less troublesome similarity relations between abstract objects (models) and reality. But what is missing in Giere's account (ironically enough given his devotion to the "cognitive approach") is a conception of how the scientist represents either models or reality. Once we put the cognitive agent back into the picture, it is easy to see that all the familiar concerns about correspondence truth recur in wondering how a scientist's representation of a similarity relation between a model and part of reality can correspond to the actual relation. Giere seems to inherit traditional problems for scientific realists and, as a bonus, the difficulty of reference to abstract entities.

Let us now ask how Giere's conversations at the IUCF might support his realism. The strategy seems to be to show that we can only make sense of the scientists whom Giere studied by invoking "entities with roughly the properties physicists themselves ascribe to protons and neutrons" (112). Now, it is not difficult for Giere to amass evidence that his subjects *believe* that they are accelerating protons —and if Giere's thesis were simply the sociological view that scientists believe in entities beyond the range of standard observation, he would have little difficulty.[2] But Giere's attempts to go further seem doomed. Why should we not explain the positioning of the neutron detectors at the IUCF in terms of the physicists' beliefs about what is going on? According to Giere, appealing to beliefs rather than to the facts about neutron trajectories would be "cumbersome" and "substantially weaker" (127). This rejoinder misses the obvious point that where the physicists put the neutron detectors depends on what they *believe* about the situation and what they *want* to achieve, as we can readily appreciate by supposing that they are unaware of what is actually occurring in nature. The very style of explanation that Giere disavows is the appropriate form for explaining behavior. *Explaining Science* does contain some passages that seek to go further, apparently posing the problem of explaining the fact that the physicists' behavior *works* (125). These passages are too brief to support

---

[2] The occasional passages in which Giere suggests that this is what he is about are outweighed by those in which he advances stronger claims. That is surely as well. Michael Devitt describes the view that scientists sometimes succeed in referring to mind-independent entities as "anti-realism with a fig-leaf"—*Realism and Truth* (Princeton: University Press, 1984), p. 15. When "realism" is reduced further to the claim that scientists sometimes believe that they are referring to mind-independent entities, the fig-leaf would seem to be transparent.

Giere's conclusions, however, since, as realists know all too well, arguments that appeal to the apparent success of science raise all sorts of complications that Giere does not consider.

If the relevance of cognitive science to Giere's realism is unclear, its *potential* significance for the treatment of decision making is obvious. But the appeal to psychological findings about failures in Bayesian reasoning needs to be handled carefully: although I agree with Giere that investigation of human cognitive practice is important (for 'ought' does imply 'can'), his treatment often seems not to focus on the disclosure of capacities and limitations but on praising the cognitive status quo. But although he appeals to psychological findings to dismiss Bayesianism, he supplies no cognitive picture for his own claim that scientists are satisficers. Distinctions here are critical, for satisficing behavior can be mimicked by Bayesians who maximize subject to constraint, Bayesian maximizing behavior can be imitated by satisficers who have an appropriate satisfaction criterion, and, in the testing situation, Bayesians and satisficers might behave just like old-fashioned hypothetico-deductivists. Indeed, so underdescribed is Giere's account of scientific decision making and so schematic his depiction of the testing situation, that I found it impossible to differentiate his own account from several familiar rivals.

The ethnographic enterprise of the final chapters seems to me to be admirable, for philosophers of science should learn more about how scientists actually behave. As a project for testing Giere's views of scientific decision making, however, it could not succeed, for, given their vagueness, those views are too loosely connected with the available data. Furthermore, the evidence from conversations with nuclear physicists and geophysicists is, though often fascinating, problematic. How much can we learn about the practice of science from interviews with a small sample of scientists? Surely more than from sitting in our armchairs, but equally surely less than from detailed studies of the behavior of all the participants in a scientific field over a period in which there is significant change.[3]

Explaining Science is a book of promises, and the promises are often unredeemed. Yet, despite my criticisms of it, I hope that it will be influential. Philosophy of science needs the cognitive approach, not because psychology has been abandoned in recent philosophy of science but because a crude version of empiricist psychology has been taken for granted. A rapprochement between philosophy and sociology is also overdue, and it is to Giere's credit that he has been able to look beyond the apparently untutored but provocative philosophy in recent sociology of science and to see the value of sociologists' depictions of scientific practice.[4] In short, this book should be read for its suggestions, for the outlining of a far more complex enterprise than Giere could have completed in the space he allowed himself. Giere's questions are, I think, far more valuable than his answers or his arguments.

---

[3] Lest it be thought that such studies are impossible, let me note that one has been carried out: David Hull's investigation of cladistic and phenetic systematics, reported in his *Science as a Process* (Chicago: University Press, 1988). Hull's study faces the potential criticism that his "field subjects" are hardly paradigm scientists—they are involved with questions that are frequently philosophical—but it is methodologically much more sophisticated than the research undertaken by Giere.

PHILIP KITCHER

University of California/San Diego

In Giere's defense, it might be urged that the laboratory studies undertaken by some sociologists of science have the same anecdotal quality. There are serious methodological issues concerning the evidential force of such studies, which deserve discussion by both philosophers and sociologists.

[4] It does seem to me, however, that Giere's accounts of the writings of sociologists of science fail to do the careful analytic work that a sympathetic philosophical reader ought to provide. In particular, his discussions of "contingency" and "negotiation" appear quite inadequate.

NOTES AND NEWS

The Austrian Ludwig Wittgenstein Society announces the Fifteenth International Wittgenstein Symposium in Kirchberg/Wechsel, Austria, August 18–25. The main theme will be "Applied Ethics and Its Foundations." The deadline for receipt of abstracts is May 31. Those planning to attend the conference should contact A. Hübner, Austrian Ludwig Wittgenstein Society, A-2280 Kirchberg/Wechsel, Markt 2, Austria.

The London School of Economics and Political Science announces the creation of a Centre for the Philosophy of the Natural and Social Sciences from October 1990. The Centre's general aim is to promote research and excellence in its field. The Centre will be run by an interdisciplinary Steering Committee comprising members of the Department of Philosophy, Logic & Scientific Method together with other LSE academics. The first director is John Worrall.

Professor and Mrs. Colin Turbayne have established a Berkeley Prize Essay competition in conjunction with the Philosophy Department at the University of Rochester. The next deadline for submitting papers is November 1. Essays should be new and unpublished and should be written in English and not exceed 5,000 words in length. Submissions will be judged by members of a review board selected by the Philosophy Dept. at the Univ. of Rochester. Submissions should be sent to: Prof. Ralf Meerbote, Chair, Phil. Dept., Univ. of Rochester, Rochester, NY 14627.

The Hume Society is pleased to announce a call for papers for its Nineteenth Hume Conference taking place at the University of Nantes, June 29–July 3, 1992. Papers may be on any aspect of Hume's writings, although the conference directors particularly encourage submissions on three themes: "Hume and French Philosophy, Past and Present," "Hume's Aesthetics," "19th and early 20th-century Interpretations of Hume." Papers should be no more than thirty minutes reading length with self-references eliminated for blind refereeing. They may be in English or French, but all papers must be accompanied by an abstract in English. Deadline for submissions is October 1, 1991. Send three copies of papers and abstracts to: Prof. Dorothy Coleman, Secretary of the Hume Society, Phil. Dept., College of William and Mary, Williamsburg, VA 23185.

Penn State Press announces a new series in Literature and Philosophy. The series will publish books in a wide range of subjects in philosophy and literature, including studies of the social and historical issues that relate these two fields. Submissions, proposals, and inquiries may be directed to A. J. Cascardi, Comparative Literature Dept., Univ. of California, Berkeley, CA 94720.

The Department of Philosophy at Bowling Green State University is pleased to announce the appointments of Edward F. McClennen as Ohio Regents Eminent Scholar in Moral and Social Philosophy, and Loren E. Lomasky as Professor.

The Department of Philosophy at Johns Hopkins University is pleased to announce the appointment of Richard Bett as Assistant Professor beginning July 1, following a term as Visiting Assistant Professor in the spring.

The Department of Philosophy at Long Island University/Brooklyn is pleased to announce the appointment of Kristana Arp as Assistant Professor.

The Department of Philosophy at the University of California/Irvine is pleased to announce the appointments of Alan Nelson as Associate Professor with tenure, and Christia Mercer as Assistant Professor.

The Department of Philosophy at the University of California/Davis is pleased to announce the appointments of Paul Teller as Professor, and Michael Dietrich as Acting Assistant Professor.

The Department of Philosophy at University of California/Riverside is pleased to announce the appointment of Georgia Warnke as Professor as of July. The Department also announces the first annual Riverside Colloquium, May 10–11, on ethics of virtue. For more information, you may write to John Martin Fischer, Phil. Dept., Univ. of Caifornia, Riverside, CA 92521.

THE JOURNAL OF PHILOSOPHY

VOLUME LXXXVIII, NO. 4, APRIL 1991

HOW EUCLIDEAN GEOMETRY HAS MISLED METAPHYSICS*

EUCLID'S geometry is so brilliant that it has always dazzled us. It shone for centuries so brightly as to blind us to any model for an established branch of knowledge but the deductive system based on a few self-evident axioms. Thus, it misled metaphysics. But that is not my theme. I want to show how it led Leibniz to some mistaken arguments in the central letters of the Leibniz-Clarke correspondence. The letters are about quite special issues in the metaphysics of space. But they have dominated philosophy since then as our models of how epistemological arguments should decide metaphysical questions. My theme is to reveal the arguments as mistakes.

Space sets ontology one of its most acute, searching, and elegant problems. To change my visual metaphor, the reason why we have not been able to see further into this subject is because a giant has been standing on our shoulders. He is Leibniz. Although the arguments in his letters are fallacious, surely he is a giant. What can make those fallacies clear to us was not available to him nor to his great opponent Newton. But first: What is the problem space sets for ontology?

I. INTRODUCTION

On the face of it, space has disconcerting properties. The list below is loosely phrased, and neither exclusive nor exhaustive. But it shows us what the problem is—on the face of things, at least.

---

* I have read versions of this paper to a number of meetings in Australia, Canada, and the United States. It owes much to the stimulating discussions it received. I am especially grateful to my colleagues at the University of California/San Diego, where I wrote the first version of it while on leave there.

0022-362X/91/8804/169–89

(i) Space is a concrete *particular,* being neither general nor abstract. It has parts not instances. There is only one space (we ordinarily think) within which everything is related to everything else.

(ii) Space stands in *concrete* relations to concrete things. I *move through* it, *see across* it to trees and cars at various distances, it *contains* me and all these concrete things. These are relations which, typically, only concrete things have to each other. Abstract things can play no part in them.

(iii) Space is *physical*—that is, space plays a prominent role in physics. Physicists are always tracing paths and trajectories, measuring lengths or constructing theoretically sophisticated, elaborate, and expensive standards of length invariance. It also figures in their theories.

(iv) Space is *not material.* As classically conceived, it has no mass, momentum, or energy. Our usual nomenclature for views of the nature of space is unhelpful. Newton denied that space is a substance just because it is not material in this sense—"it does not support characteristic affections of that sort that denominate substance, namely actions such as are . . . motions in a body."[1] That seems right. If so, then the view that it acts, is material, or has mass or energy surely ought to be called substantivalism. That is the right way to describe space-time as we find it in General Relativity.

(v) Space is *not perceptible.* We never see space or its parts; we neither hear them, smell them, nor touch them. Yet we can see across it. We can see at a glance that just two objects are now visible though we "can't tell them apart." Suppose that they are in fact qualitatively identical. That surely means that we can see immediately that the two are in different places even though we see only the things and not the places.

The acute problem for ontology lies in the sharp clash between the first three properties on the list and the last two. The first three properties seem to tell us that space is something; the last two that it is nothing. This problem about space in our ontology is unresolved. *Realism* is my name for the view that space is a thing, earns a place in our ontology, and has (something like) properties mentioned in the list. I distinguish realism from *substantivalism.*

The precision and power of geometry make the space problem elegant because the arguing can be so neat, clear, and determinate. The curious relations of space to perception (feature (v)) demand that we frame with particular care and rigor the epistemological principles on which we either reject or accept an ontic commitment to space. So the problem is searching, too.

---

[1] For a corrected translation of a part of the *Unpublished Scientific Papers of Isaac Newton,* A. Rupert Hall & Marie Boas, eds. (New York: Cambridge, 1962); see Howard Stein, "On Metaphysics and Method in Newton" (forthcoming).

II. DOUBLING IN SIZE

Let us begin with a familiar and graphic challenge. It snugly fits the mould of Leibniz's arguments, though we owe the first form of it rather to Henri Poincaré. Both the style and the mistake are the same as in Leibniz's letters. Every philosophy student surely hears this argument at an early stage of his education. It is one of our best-known intuition pumps. It will be useful to start with it.

Suppose everything were to double in size overnight. Nothing would betray the change. Round pegs would fit snugly into their old round holes and other holes squarely resist their entry as before. Everything would be related just as it was to everything else. The difference would really be no difference, the supposed change no change at all. Nocturnal expansion is absurd.

In more detail, the argument has to run like this. At first glance, space looks a useful posit: the size of a thing is the quantity of space it fills, its shape is the disposition of the thing's parts in space, the spatial relations of one thing to others is fixed by their places in space and the lengths and directions of the spatial paths that join one place to others. The spatial relations between things look as if they are all mediated by relations of things to space.

The doubling argument seeks to cut this tie. It is a kind of reductio argument in which the absurdity is not a contradiction but an idleness. It is a thought experiment in which the realist gets his space but finds that he can do nothing with it. It shows how to change the thing-space spatial relations of every thing yet retain all thing-thing spatial relations. Conversely, the thing-thing spatial relations are shown not to determine any thing-space spatial relations. The relationist's gift of space is revealed as ironic—it gives nothing at all. There are no thing-space relations. So it is needless to posit an entity that exists only to found them. Space is a chimera.

All this rests on some kind of indiscernibility or inconsequentiality principle. It might be this one:

> IP: "Inconsequential properties" are not properties.

Here 'inconsequential property' means something like "property which can change without any accompanying change in specifiably privileged properties." For the case in hand, we have two ways of saying what is privileged. Observable properties are privileged; more specifically (and less naively), thing-thing spatial relations are privileged. I shall stick to the latter account which will do for Leibniz himself. So let us distinguish thing-thing spatial relations from thing-space spatial relations. Thus, we can state the detachment thesis:

Detachment thesis: thing-thing spatial relations are logically independent of thing-space spatial relations.

This will put us in a position to deny the realism of the introduction. All thing-space spatial relations are idle; so the relatum of these relations is idle, too. Space is no part of our ontology. Only the concrete things that stand in spatial relations together with those relations themselves are real; space is an unwanted absurdity.

The detachment thesis lies at the foundations of a prominent reductive theory of metaphysics: spatial relationism. Spatial relationism can fairly promise to reduce space to thing-thing spatial relations as long as these relations can be detached from thing-space spatial relations. The detachment thesis clears relationism of the otherwise plausible charge that it is circular. The detachment thesis is no help in showing how the relationist can complete his programme. It simply legitimates his title to use thing-thing spatial relations in the reduction. It offers an escape from circularity. It is not obvious how to escape circularity without it.

III. OBJECTIONS TO THE DOUBLING ARGUMENT

Perhaps the reason why Leibniz never used the doubling argument is that he read the first of Galileo's *Two New Sciences*. Galileo points out that doubling a thing in all its lengths (scaling it up) will quadruple its cross-section areas and multiply its volume and thus its mass by eight. The load-bearing factors of the object are essentially functions of area, but the loads they must bear for the thing to be stable are functions of its mass. So scaled-up models of strong small things, like ants, may be unstable, as producers of gigantism horror movies often find. The bigger they are the harder they stand—or fall. What, too, about the speed of light, and other constants—Planck's for one?

Poincaré thought that he could avoid this objection and it is evident that Hans Reichenbach, Adolf Grünbaum, and many others thought so, too. This was challenged by George Schlesinger.[2] I shall not try to umpire that debate, though it surfaces again briefly in section VI, objection 2. My main objection, in the next paragraph, is simpler and independent of this. I think it is more incisive.

The claim that doubling has no consequences rests on a quite precise and unique assumption about thing-space relations. Things must be doubled *in Euclidean space*. Among the geometries of constant curvature, only Euclid's geometry allows similar figures of different sizes. If everything doubled in size overnight in a non-Euclidean space, everything would change in shape. Thing-thing spatial relations would change with their changing thing-space relations. Further, in non-Euclidean spaces, doubling in size would differ observably from tripling in size in non-Euclidean spaces—it induces weaker changes in shape than tripling in size. The geometrical facts are familiar enough. But their significance for metaphysics has gone unnoticed. Before I turn to that, I shall illustrate the geometrical facts as simply as I can in the next, starred paragraphs. The reader who does not need the illustration can skip them.

---

[2] "What Does the Denial of Absolute Space Mean?" *Australasian Journal of Philosophy*, XLV (1967): 44–60. And cf. Adolf Grünbaum, "The Denial of Absolute Space and the Hypothesis of Universal Nocturnal Expansion: A Rejoinder to George Schlesinger," *Australasian Journal of Philosophy*, XLV (1967): 61–91.

* It is easy to see what doubling in size comes to in two-dimensional spaces (surfaces) that do not have the character of the Euclidean plane. The consequences generalize fairly intuitively to their counterparts in three dimensions. Let us begin with the surface of a sphere. A triangle in that surface will be any three-sided figure composed of sides that are geodesics of the sphere, that is, great circles of it. If the triangle is small relative to the radius of curvature of the space, then the sum of its three angles will be close to a straight angle. But a big triangle will have no such shape. Imagine, as one side of a large triangle, an arc of a great circle just less than half the circumference long. At its end points, construct two more great circles to meet at a point polar to this first side of the triangle. The angles at each of the vertices on the equator will be a right angle (half a straight angle), so together they will sum to a straight angle. The angle at the pole will be close to a straight angle, too, so that the whole angle sum will be slightly less than 360°—nearly a complete revolution. That is a gross (but not the grossest) departure from Euclidicity. This defect (as it is called) is a function of the ratio between the area of the triangle and the radius of the curvature of the space. So, if a triangle on the sphere is doubled in size overnight, it changes shape. Not only is its shape different; it now has a shape *it could not have had* yesterday.

* Spaces of positive constant curvature, of which the spherical surface is a two-dimensional example, are finite spaces. To double everything seems to mean doubling the size of configurations of objects (e.g., stellar triangles) too. But then there might not be enough space for them to double in. If stars are spread more or less uniformly all through a finite space, the experiment is not thinkable. And if the configurations are not doubled, the doubling just of star sizes will obviously change thing-thing spatial relations. Some spaces of constant zero curvature are finite, too (the torus, for instance). So, even when the local character of space is everywhere Euclidean and doubling is locally indiscernible, we cannot guarantee that it will be globally indiscernible. The operation cannot be imagined as performable no matter what geometry is in question. We have to know what the space we are dealing with is like.

* We get defects from doubling lengths in negatively curved spaces, too, though here the angle sums get smaller as the lengths increase.

* If triangles change in shape on doubling in these spaces, then so does every other shape. Any polygon can be decomposed into triangles;

any closed curve can be approximated by some polygon. As surfaces change in two dimensions, so do the volumes they bound in three dimensions; and so on up. The result is perfectly general.

* If the space varies in its curvature from point to point, then the consequences of doubling will depend on where the doubling begins and how it spreads. Imagine, first, a surface that varies its curvature from point to point: it will resemble a range of smooth hills and valleys. In some places, the curvature will be positive (hill tops and valley bottoms); in other places (saddle backs or hill sides), the curvature will be negative and, thus, since we suppose the curvature varies smoothly, some places in between will have zero curvature. Now, it will certainly make a difference to thing-thing spatial relations from which center (whether occupied or empty) the doubling occurs. If one point rather than another is the center, then some things will be moved further than others by the expansion and that, in turn, will determine whether they themselves are doubled in regions of positive, negative, or zero curvature. So there will be uncountably many ways of performing the thought experiment and uncountably many distinct thing-thing relations resulting from them. So doubling-in-thought experiments depend on which thing-place relations of spatial occupancy (position, place) are preserved, if any, and how others are altered.

* Perhaps it is worth pointing out that such changes are in general quite grossly observable. How are we to measure the angle sums of large triangles? Here is a crude method. At each angle, use a stretched string to draw a circle of standard radius by anchoring one end of the string at the vertex. Now lay the string along the drawn circle with one end of the string at the point where one side of the triangle cuts the circle and mark along the string where the other side cuts the circle. So we measure the length of the arc of a unit circle that coincides with the rotation through that angle. (The length of the chord would serve equally well provided we note when the angle measured by it is greater than a straight angle.) We can transfer the results of this operation from one vertex to another and add the angles simply by adding the arc lengths successively round the unit circle. The one piece of string will give us both the unit circle and the arc lengths. Now, if we carry out this operation of adding the two distant angles of the large triangle to the nearby one, we shall find that the arcs add to almost a complete revolution (or more for still larger triangles). There is nothing subtle about the defects.

The Leibniz-style Poincaré objection to the realist view of space is fallacious. It is an argument to the detachment thesis—that thing-thing spatial relations are independent of thing-space spatial relations. But the crucial invariance of thing-thing spatial relations requires attachment, that is, precise and unique thing-space spatial relations. Things are to be expanded *through Euclidean space*. Otherwise, it will not even appear to legitimate its claims to inconsequentiality. What it presupposes is inconsistent with what it concludes. Note in particular that the dependence of the outcome of doubling on the nature of the space applies with equal force to the special and actually indiscernible case of doubling in Euclidean space. It is not only when doubling will change shapes that we need to specify in which space we are performing the thought experiment. That the doubling operation in Euclidean space has no thing-thing consequences depends on specific symmetries unique to the structure of that space. These symmetries pick out the similarity transformations from the rest. They do not show that there is no structure and no space. They mean that there are specific structures in a specific space.

The string of words 'A is double the size' is not a sentence; it is not syntactically complete. If you like, it is a chimera. To have a thought, true or false, we have to know what the thing is twice as large *as*. Just so, the "thought experiment" given by the directive 'Double things in size overnight' must be syntactically complete. But now it is not enough to know that things are to be double the size they were. We need to know in (through) what they are to be doubled. In each case, the chimera is a monster—ill-formed as a sentence or a thought. The imperative that directs us to perform the operation is syntactically incoherent. The "thought experiment" is unintelligible. We have to specify the space. (Of course, we shall get general sense out of the idea of doubling things in *some* space, since that is syntactically coherent. But like the idea of painting a thing some color, it tells us little about how things will look when the job is done.)

Further, we have to double things in space (expand them through it) to hope to support a detachment thesis: else it will no longer be a thought experiment that detaches thing-thing from thing-space spatial relations. The relationist has to begin by giving the realist a space. His problem is how to get this gift back again.

IV. LEIBNIZ AND THE DETACHMENT THESIS

Leibniz brought several arguments against Clarke to "confute the fancy" of those who posit the chimera of space. All of them founder on the same rock and by much the same style of navigation as we have seen. The arguments need symmetries peculiar to Euclidean space. The appeal fails without them. (Leibniz ties his arguments to a theological dispute which is beside the point of this paper. Perhaps theological issues warp any straight relevance to modern issues. In that case, what I have to say will concern interpretations of his arguments in the more recent philosophy of science.)

Let us begin with the most important of these, the relativity of uniform motion in Newtonian mechanics.[3] How does this engage the inconsequentiality principle? To every body in the universe, whatever its state of rest or motion, add in thought the same velocity vector—the same speed in the same direction. Then everything will change its thing-space spatial relations; everything will be moving through space. Leibniz claims that this will make no difference to thing-thing spatial relations.

Newton's mechanics entails Euclidean space. It presupposes that we *can* add the same velocity vector to every thing. We can do that only if each added vector is parallel to every other. But the existence of parallels is peculiar to Euclidean geometry.

Consider a cloud of small particles that move independently (do not exert significant forces on each other). It is dust. Then, like its component particles, the cloud itself can move at uniform velocity through Euclidean space, free from the action of any force. It will change its thing-space spatial relations (position) continuously but the relative position vectors among the particles in it will not change; they are the vectors that relate particles in the cloud to each other at any time by their distances and directions. So the cloud does not change its shape as it moves. In motion, it is just as if the cloud were at rest. That, in turn, is possible because every particle in the cloud can have the same velocity vector. But the idea of the same velocity vector makes sense only if the velocity vector on any one particle is parallel to that on any other. Mechanics allows a symmetry among the relative uniform motions of point particles. That is Galilean relativity. But we need Euclidean geometry to extend this to moving volumes. In other geometries of constant curvature, there are no parallels. This may be because all geodesics intersect somewhere (positive curvature). It may be because, given any geodesic *G* and a point *P* not on *G*, there will be infinitely many other geodesics through *P*, coplanar with *G* and *P*, and which do not intersect *G* anywhere. That is the case when the curvature is negative. Nevertheless, there will be no geodesic in the plane and through the point that is everywhere the same distance from *G*. It is often said of spaces of constant negative curvature that they have many parallels. This is not true in the sense needed for the thought experiment if it is to yield the detachment thesis.[4]

---

[3] Leibniz's Fourth Paper, §13; Fifth Paper, *The Leibniz-Clarke Correspondence*, H. G. Alexander, ed. (New York: Barnes & Noble, 1956), pp. 38, 73–5.

[4] For a more detailed account of what the motion of dust clouds and elastic solids might be in spaces of constant nonzero curvature, see my "What Can Geometry Explain?" *British Journal of the Philosophy of Science*, xxx (1979): 69–83.

If a dust cloud moving in non-Euclidean space changes shape, that is not an action of space on things; no more so than the constancy of shape of a moving dust cloud in Euclidean space is an action of space. If the motion of each particle is uniform (as I have been imagining), each will move along a geodesic. That divergence of geodesics in non-Euclidean space is not something that the space does to things that move along them. Space is not a cause, and the explanation of the changing shape of the cloud in motion is geometrical not causal.

Leibniz claims that a uniform motion of simply everything would be without consequences. But this depends on the objects being attached to (being moved through) a quite precise and unique kind of space. Our thought experiment does not show that we can freely detach thing-thing spatial relations from thing-space spatial relations. The relations depend on each other. The argument to the thesis is fallacious; it presupposes (tacitly entails) the attachment it attempts to deny. Here again, quite clearly, the attachment is presupposed not just in the case of non-Euclidean spaces where uniform motion is well-defined and observable. The inconsequentiality Leibniz needs is a symmetry of Euclidean space. He needs the symmetry to be concretely there in something real and not chimerical.

This suggests that Leibniz simply assumed that Euclidean symmetries were inescapable. But it is not quite so simple. He seems to have thought he could derive the symmetries just from the indiscernibility of spatial points one from another. His "proof" runs as follows.

> Space is something absolutely uniform; and, without the things placed in it, one point of space does not absolutely differ in any respect whatsoever from another point of space. Now from hence it follows (supposing space to be something in itself, besides the order of bodies among themselves) that 'tis impossible there should be a reason why God, preserving the same situation of bodies among themselves, should have placed them in space after one particular manner, and not otherwise (Fifth paper §5).

The passage does not make it clear whether Leibniz thinks the uniformity of space derives from the indiscernibility of points or vice versa. But we should not accuse him of taking the symmetries for granted simply from Euclid's geometry.

In any event, his conclusion does not follow. Spatial structures richer than topology are not gained by adding to properties of points. Affine and metric structures are properties of the space as a whole and "emergent" relative to properties intrinsic to points. The presence, absence, or variation of curvature is not an intrinsic difference among points. It depends on the structure of the space. Curvature is a limiting property of points, deriving from the (possibly varying) structure of the space round them.

Leibniz also tells us that it could make no difference to thing-thing spatial relations if everything were displaced some arbitrary distance and all in the same direction. As before, the phrase 'all in the same direction' requires parallels. But there are new problems. He tells us that there would be no change to thing-thing spatial relations if things were interchanged East and West.[5] (Whether this is supposed to be the result of a reflection, or a rotation round an axis makes no odds.) These displacements are not to be thought of as processes of moving things to new places. Leibniz's theological metaphors make the point clear. It would make no difference had God created things not where he did, but so that each came into being 100 meters further North or with East and West interchanged. (Put aside proper worries about North, East, and West as the same directions everywhere.) The point is that the displacement would yield differences if the curvature of the space varies from point to point. Indeed, if our thought experiment asks us to imagine perfectly rigid bodies, not even God will be able to perform it. A rigid body cannot be relocated in a region where the curvature is different, since no shape in a region of one curvature is the same as any shape in a region of different curvature. An elastic body might be relocated by distorting it, but a *perfectly rigid* body cannot be distorted. God cannot act a contradiction. Among elastic bodies, we can expect rubber objects to distort more than steel bodies and expect that it would be easier to relocate them. But we can detect and measure these various distortions. We are not dealing here with Reichenbach's universal forces, but with common or garden elasticity, tensile strength, and the like.

The inconsequentialities of Euclidean space are not the product of a confused hypothesis about a nonentity. They are the symmetries of a unique and well-defined entity with a perfectly definite, if bland, structure. Change the structure of that entity and you have the thing-thing relation consequences that Leibniz thinks impossible. We cannot argue as he does with the inconsequentiality thesis as premise. It is false.

V. SOME CONCLUSIONS SO FAR

The detachment thesis lacks support. Any version of the reductive relationist program for banishing space from our ontology which depends on Leibnizian arguments for the detachment thesis is circular. The relationist thinks that his hardest job lies in completing the reduction. But that is really the easy part. The relationist uses what seem at first glance to be parts of space (such as the paths across which two things are at a distance or volumes of space in which the matter of things is disposed). If he can show his entitlement to use them without circularity then his residual problem looks soluble. It is the problem how to determine the character of a whole by specifying relations among certain parts of it. It is by no means a contemptibly easy problem. But it is not obviously a metaphysical one either. The vexed question lies elsewhere in the use of the relations for reductive purposes. Can we detach thing-thing spatial relations from thing-space ones quite straightforwardly? The prominence of Leibniz-style arguments in the literature suggests that, on the face of it, we cannot; these arguments aim to show something that needs to be shown. Suppose that it cannot be shown. Then, as ontologists, we should be no less worried by the nature of the parts of space (volumes, paths, points) than by the nature of the whole. Whole and parts equally have the same awkward features listed in section I. The real problem is to show that thing-thing spatial relations do not depend on thing-space relations. Relationism without a detachment thesis to establish its freedom from circularity lacks reductive teeth.

---

[5] Leibniz, Third Paper, §5; Alexander, ed., p. 26.

Leibniz appeals to the "indiscernibility principle"; it is dubious. Its converse is highly plausible, however. We might call it the "discernibility principle" and express it as follows:

DP: Consequential properties are real properties.

This is the contrapositive of the converse of IP. We can gloss 'consequential property' after the style of 'inconsequential property'. Suppose we adopt DP. Then we have really a rather plausible argument that position in space and the quantity of space filled or occupied by a material thing are real relations. So some thing-space relations are genuine and not chimerical. In general, change in these relations changes thing-thing spatial relations in consequence. If the latter changes are real, so are the changing properties that induce the consequences. But then the thing that has the properties must be real, too.

Thirdly, and more generally (but more vaguely), Leibniz's conclusion has been our paradigm for a whole range of epistemological claims about the probity and point of concepts in science and outside it. The path for many of these has been the easier because of the apparently brilliant success of their Leibnizian forerunners. Reductive epistemology did not begin with Leibniz, but his incisive attack on Newton wrote a distinctive new page. This, together with the fact that classical physics after Newton seemed to endorse his attack, lent a respectability to reductive empiricism, in one form or other, which it would hardly have gained otherwise.

Theses that something or other is unobservable in principle have proved vulnerable. Most of them were proposed as having a status unhappily like that of synthetic necessary truths, though, like the verification principle, used to deny that there could be any such truths. They are as fragile as allegedly synthetic necessary truths deserve to be. Yet they have motivated various draconian principles by means of which seemingly respectable conceptual structures were run out of town. This vigilante epistemology has seemed judicious largely because Leibniz's dazzling arguments about space supported it. If Leibniz's arguments are our paradigm for the probity of inconsequentiality and reductivism, their failure might become our paradigm for impeaching them.

Thus, the role of epistemology in metaphysics has been regularly misunderstood and overstated by confining ontology to a dependent and sharply limited role in a puritanically conceived theory of knowledge. In this mistaken and destructive idea of things, a fixation on Euclidean geometry played no small part. It is ironical to reflect that it is the very simplicity and symmetry of Euclidean geometry which has misled us. The simplest possible spatial world is confusing in its very simplicity and blandness.

VI. FURTHER ASSUMPTIONS AND CONSEQUENCES

Ontologically, Euclidean space is surely just the same kind of thing as non-Euclidean space. Newton's theory demands that we include Euclidean space in our ontology. The metaphysical nature of this commitment is no more disturbing than when our theory commits us to a more readily observable non-Euclidean space. No doubt Euclidean space is bland. But that cannot change the nature of an ontic commitment to space. Non-Euclidean space has all the disconcerting features in our list of section I just as surely as Euclidean space has them. So I make the following assumption:

> A1: Structural symmetries do not affect the metaphysical nature of an entity.

The structural symmetries of the entity may well alter the way it enters into our theories, however. We now see that it is the realist who can postulate spatial symmetries and thus explain why, in Euclidean space, we can never observe the difference between doubling and constant size, between rest and uniform motion, and so on. The realist owns the symmetries, not the relationist. It is realist's realism that tells him (and tells us) why we cannot care about the difference between one of these states and another. In fact, it is the relationist who has no clear, coherent, noncircular account of the relativity of motion. That relativity rests, not on nothing, but on symmetries of a real entity, Euclidean space. *Therefore, the realist does not have to answer the objection that is standardly brought against him.* The objection is that the realist is obliged to make these inconsequential distinctions and must take them seriously, whereas the relationist need not. But, on the contrary, the realist is just the person who need not concern himself with them, since he and he alone has a clear picture of their source and their nature and their lack of consequences. He alone has a coherent account of when they hold and how they may fail. They spring from the character of Euclidean space. How can anyone oblige him to care about what *the nature of his postulate* makes plain to him that he *cannot* care about?

Let us use *absolutism* to describe the view that there are some symmetry transcending properties that matter for science or for metaphysics. Then realism does not entail absolutism. Rejecting absolutism takes no step toward relationism. Newton was an absolutist.

Nonetheless, motion is motion-through-space. '*x* moves' is good grammar, for we can say whether it is true or false if it is contextually clear what the tacit relatum is. But the deeper syntax of sentences of motion requires that the referent be filled out on demand—the demand of how the sentence is to be understood. Now, the non-Euclidean cases make it plain, first, that the demand *need* not be met by giving an observable thing as the tacit relatum. We may say instead whether or not things move in a non-Euclidean space. But that is not all. Familiar motion, relative to other things in space, does not yield an equally significant idea. It is a weak or recessive sense, bound to yield to the deeper more general requirement that a thing is in motion through space. Thing-thing relative motion might turn out to be a symmetric relation in a suitable theory. But the deeper thing-space motion is asymmetrical on semantic grounds. Space does not move relative to things that it contains (but see section VII, objection 1).

The realist can draw a clear map of the path followed by the post-Newtonian classical physicists. He can relate rest and motion to any of the frames of reference that dynamics and the symmetries of Euclidean geometry together select as indistinguishable. It is at least as clear a map as a relationist can draw. It does nothing to undercut his commitment to a space with indiscernible symmetries. The relativity of motion rests on the commitment. Every frame of reference defines a Euclidean space with the metaphysical properties noted in our list in section I. We may, if we wish, consider these as different hypotheses about what is at rest, what in uniform motion, and the like. Our realism tells us that the difference between these hypotheses cannot matter. There is no clear position that allows us to say that space forms no part of the ontology of post-Newtonian physics. That ontic commitment is there in every choice of frame. It is unclear how relationism avoids begging this question.

I conclude that realism does not entail absolutism.

There is a wild card in this deck, however. It is the metaphysics of continuants. Continuants have spatial parts but not temporal ones. They have histories, but the history of a continuant is a property of it, not a part. Suppose that you and I are continuants. Then I am the person who wrote the paper you are reading: there is no paper-writing part of me nor paper-reading part of you. Neither of us is extended in time. Our histories are our tensed properties, not parts of us. Newton supposed that space was a continuant (though not a substance) and, for this reason, was obliged in general to admit that questions of the identity of places at different times were meaningful and admit that there were different indistinguishable hypotheses to choose (freely) among. A realist need not follow this path. Instead, place classical mechanics in space-time; this will mean that questions about which is the same place at different times cannot arise because space-time is not a continuant entity. Nothing in this bows to detachment theses or spatial relationism: space-time causes much the same consternation among ontologists as space causes and for much the same reasons.

But neither that nor the fact that a realist sees he sometimes cannot care about distinctions between rest and motion means that the questions about rest and uniform motion through space are meaningless. They are idle in the case where space is Euclidean, but only just. Give the space a little uniform curvature and the distinctions engage at once with thing-thing relations. Surely the limiting case of zero curvature inherits the sense of the range of cases of which it is the limit? It is not clear why not. Suppose a space that is changing its curvature from (everywhere constant) positive to (everywhere constant) negative curvature. Do the questions not retain both their point and their meaning as the changing space passes through the limit between the two? I assume that it does.

Suppose we deny the assumption A1. Instead we assert:

A2: We cannot be ontically committed to a space with symmetries.

It would seem to be IP, the indiscernibility or inconsequentiality principle, that might motivate A2. But if that is right, then it would seem to commit us to the view that, though there might well be non-Euclidean space (best, perhaps, space of variable curvature), there could never be a Euclidean space. It is too symmetrical, observationally bland, or anaemic to exist. J. L. Aústin once spoke of the idea that existing is like breathing, only quieter. We seem to encounter something like this here, but without the quietism.

Perhaps it is this view which John Earman[6] recommends to a reconstructed ahistorical Leibniz in almost the only place in the philosophical literature where the ideas I have been raising seem to have been mentioned. The relevant two or three sentences come in what appears as a casual afterthought in his splendid paper. Earman points out that Leibniz's arguments fail if the space lacks the relevant Euclidean symmetries. He grants Leibniz anachronistic knowledge of the variety of geometries, and imagines him objecting that Newton used too bland a space to allow him to claim its real existence. Or so I understand Earman. The brief passage is acknowledged and repeated, almost verbatim, by Michael Friedman.[7] This view of things seems to portray Leibniz as insisting on A2 and going on to claim that Euclidean space could not exist, though a non-Euclidean space might.

But I find this suggestion unacceptable. It confuses geometrical differences with ontological differences, or blandness with nonexistence. If Leibniz had indeed raised this objection, surely a decisive reply would have been available to Newton. For, first, neither of them had the least *empirical* grounds for thinking that a non-Euclidean geometry described the structure of space; second, Newton would rightly have claimed that Leibniz was attempting to saddle him with a needlessly complex geometry in order to satisfy an obscure maxim of epistemology which has no place in natural philosophy. Surely the conclusion that there could be non-Euclidean spaces but could not be a Euclidean one is absurd (though not contradictory). A2 is the child of IP, the inconsequentiality thesis; it leads to absurd consequences, so any version of IP that entails it must be false.

Further Leibniz's argument is fallacious. It is not just a matter of the falsity of his tacit premiss—that every possible space has symmetries of Euclidean space. The light shed by non-Euclidean geometries merely throws into sharp relief an already present appeal to a symmetrical space. Leibniz appeals to that symmetry throughout and it would nonetheless belong to an entity of the sort he wants to deny, but has to presuppose, even if every possible space possessed it.

---

[6] "Who's Afraid of Absolute Space?" *Australasian Journal of Philosophy*, XLVIII (1970): 287–319; esp. p. 303.

[7] *The Foundations of Space-Time Theories* (Princeton: University Press, 1983), p. 221.

Finally, I assume that realism is ill-described as substantivalism. Newton said that he was not a substantivalist and gave a clear reason for saying so. Space has no action on things. Another idea of a substance is simply that of a continuant. Neither of these is entailed by realism but only by specific forms it might take. In General Relativity, mass and energy are features of the space-time metric and affinity. I find it strange that this is described as a relationist view of space and opposed to what is called substantivalism. What have relations to do with it? What else would one want to describe as substantivalist other than the opinion that space-time stores energy and mass, that particles may be generated out of and decay into space-time structure? Certainly not the view that *is* described as substantivalism—that the manifold, despite being undetermined in its structure by the matter tensor, is real. The manifold is too attenuated to be cast as a substance. But it is nevertheless arguably real and the view that it is is surely best so described—it is realism, not substantivalism. Of course, any substantivalist view will make the distinction between container and contained less straightforward than Newton supposed. But that is no feather in the cap of traditional relationism.[8]

VII. REPLIES TO SOME OBJECTIONS

1. Let us reconsider nocturnal expansion. When we performed the thought experiment, we did not double strictly everything. We did not double the space. But if we double the space as well as the objects, then there will be no change in thing-thing spatial relations. So is not this Leibnizian argument valid after all?

No. The argument has to yield a particular conclusion, the detachment thesis. It is that thing-thing spatial relations are independent of thing-space spatial relations. But the revised thought experiment does not detach thing-space spatial relations from thing-thing ones. It preserves them. Since it is not a thought experiment in which there is a detachment, it is powerless to defend a detachment thesis.

Perhaps this is a superficial reply, because we have given no meaning to the operation of doubling the size of space. We have yet to give it even a decent syntax—make a well-formed description of it.

---

[8] See John Earman and John Norton, "What Price Substantivalism? The Hole Story," *British Journal of the Philosophy of Science*, XXXVIII (1987): 515–25.

We have not said through what the space is to be doubled. Before we can give even the meager "sense" of well-formedness to doubling our space as well as the things it contains, we must suppose that there is some higher dimensional space that contains it. But a higher dimensional containing space for our space is just an idle fancy. It plays no role in Newtonian mechanics or of any other actual theory of ours. To give to doubling the size of space something we could fairly call a semantic meaning (something stronger than well-formedness) the higher dimensional space has to be given some point in the theory within which discussion is going on. We have yet to see that point. So the idea of doubling space as well as things lacks meaning —so far.

But even if we made the fancy less idle by giving some point to the higher dimensional space, it does not follow that doubling our space through this higher dimensional space would be inconsequential. That depends on whether the wider space is Euclidean. One obvious way to give speculative point to the wider space is to make it non-Euclidean, for then the doubling of our space would change both its shape and that of the objects it contains. The doubling would be consequential. We are still fixated on Euclidean space if we suppose that doubling the space will change no thing-thing spatial relations.

It might seem that the objection has no need of this further space, as my colleague Michael Simpson pointed out to me. The curvature of the space is intrinsic to it, not dependent on some higher dimensional containing space. So it seems that we might double the radius of curvature (i.e., halve the curvature) without referring beyond the space to another which contains it. Then the objection makes sense. We can all agree, however, that metrics equivalent up to a scale factor are not meaningfully different even when the space is finite because of its curvature. This may look like an agreement among relationists and not something to which I can happily turn here. But the point is syntax (addicity) as before: a scale factor can be given sense only when we have something to which to link our space by such a relation as 'Six times as large as . . .'. We can make sense of something being as large as it was only by reference to the quantity of space it filled. As I see it, a relationist ought to insist on this just as firmly as a realist does. The further debate is whether anything could fill the bill as a relatum unless it is somehow privileged (observable, changes thing-thing relations, etc.). We all agree on a point about well-formedness: doubling the space makes no sense as an intrinsic nonrelative operation. The dispute is about whether and how we might give relative sense to it. If the hypothesis of a containing non-Euclidean space through which our space is doubled has consequences for thing-thing relations, the relationist will have difficulty in disallowing that it is meaningful. I am suggesting that whoever raises the specter of somehow doubling the space is, once again, giving me what he hopes is the empty hypothesis of a containing space and defying me to make use of it. But the use is clear.

It is much the same if our Leibnizian fancies fly to the bizarre height of adding to the space of our objects the same velocity vector we added to the objects themselves. Now, when everything moves at a uniform velocity, our space moves, too, along with the objects. The fancy is well-formed only if space-and-objects, holus-bolus, moves through a wider containing space. But, again, if this wider space is non-Euclidean, then both our space and the objects in it will change their shapes, for they will have no parallels in the wider space for the points to move along. The appeal to space as among the things to be changed by any of these thought experiments gets us no closer to the inconsequentiality Leibniz needs.

Contrast this idle fancy with something else. General Relativity is a theory in which tensor equations describe the physical fields. The fields are determined by quantities that might roughly be described as observables. They fix the metric, the affinity, and other things. But they do not determine the lie of these tensor fields on the basic manifold. They do not even fix how the fields lie in local regions beyond which everything is specified. That is the essential thrust of the hole argument. Nevertheless, the manifold is not like the idle fancies just discussed. We cannot define tensor fields except on manifolds: books on tensor analysis begin with manifold theory, as a casual look at chapter headings will show. The same is true of many books on General Relativity. So we cannot begin a tensor theory without first postulating a manifold. General Relativity leaves us with a manifold that has symmetries permitting a variety of different placements of the tensor fields within it. The symmetries show us that it cannot matter which lie on the manifold we choose. The results in terms of observables will not change. But this does not mean that we can take our postulate back and give up realism about manifolds. For, here again, it is in the structure of the manifold—from which the theory must begin—that the symmetries lie. They cannot lie in a nonentity. Nonexistence has no character—in particular, not a symmetrical character. The hole argument fails as an objection to realism. But, again, the realist's perception of the symmetries in the entity he postulates leaves him perfectly free to care nothing about these differences. He plainly sees from the nature of what he postulates that he cannot care about them.

This point is worth making because it may help us to formulate more useful principles about what is meaningful and, rather differently, what to care about in scientific or any other sort of theories. This is a really difficult task. The IP is visibly too strong. Yet we do not want to admit as meaningful every sentence containing some terms semantically anchored in their theoretical role. The preceding example of doubling physical space relative to an idle space meets that weak condition. 'Space' is a semantically anchored term. The idle space is a space in the same sense as physical space, so the sentence about doubling space is syntactically well-formed and thus it is weakly in order semantically. Until we give the idle space a role in theory, we ought to deem the doubling-of-space sentence meaningless, though open to acquiring a meaning. For who can say how theory will advance next? The case of the manifold is different: it certainly has a role in General Relativity. Let those who deny it show us how to get along literally without it instead of merely pointing out its symmetries. But I shall content myself here just with the easier task of concluding that there is much work still to be done in this regard.

2. Here is another objection. Nothing said so far takes us beyond the observation of thing-thing spatial relations. These are all we have under the hypothesis of non-Euclidean space to show us that things would change if doubled over night. Even then, the observations are not enough to establish that the space is non-Euclidean. Poincaré, Reichenbach, Grünbaum, and others have shown with ample clarity that no observations of the sort envisaged can suffice to commit us to a space with any special character whatever.

Reply: Our central question is the argument for the detachment thesis. The issue concerns this conditional: if we change thing-space relations, there are no consequential changes for thing-thing relations. This conditional thesis has been refuted. The objection challenges us to defeat the converse conditional: if thing-thing relations change as envisaged, then they are not consequences of new thing-space relations. But we have no need to refute this. It is not the detachment thesis.

Yet we could refute it. At least, we could argue against it in perfectly familiar ways. The objection throws gratuitous obstacles in the path of reasoning from the evidence to the best explanation. That sort of problem is in no way peculiar to arriving at good geometrical explanation. That is a perfectly general problem for scientific theorizing. We are too familiar (from W. V. Quine and Pierre Duhem) with the way in which tinkering with auxiliary hypotheses and initial conditions can make any body of evidence consistent with ugly, arbitrary, deviant theoretical conclusions. Although the metaphysics of space has proved a happy playground for the invention of obstacle courses, there is no reason to suppose it suffers in a more intimate way from the general malaise. The DP together with the nature of non-Euclidean geometry suggests that the hypothesis of space can do genuine explanatory work (even though it makes no appeal to the action of space). What we have to do here is to argue for the best explanation of the phenomena. Although such arguments are not formally valid, they convince rational people.

Given the evidence we have been speculating about, inference to a space with one or other definite geometry looks like the most elegant and fruitful explanatory hypothesis. We know how to treat space (or space-time) as something that has concrete relations to concrete things and how these changing relations might explain observable changes in the world. General Relativity does this for space-time. But geometrical explanation of concrete changes by means of concrete spatial relations is clear, too.[9] As far as I can see, the main reason why we hesitate to accept geometrical explanations in these cases is that we think the idea of something immaterial and imperceptible—space—is absurd. This does not look like a strong reason. The idea of the material has undergone drastic changes in physics over the last century; the idea of what is observable is anthropocentric and local. It has usually been held that the postulate of a space with a well-defined geometry can do no explanatory work. But it plainly can. Our sense of uneasiness about it begins to look like a vulgar human prejudice.

Is it coherent, meaningful, or intelligible to commit oneself ontologically to an entity that has the sort of stressful combination of properties listed in the introduction? Surely the answer to that modest question has to be 'yes'. It is hard to see what sort of objection can be seriously raised against the intelligibility of entities like space, unless it is that space is neither perceptible nor material. Of course, spatial kinds of things are not perceptible and generally not material. But it is not clear why these should be regarded as defeating the commitment as long as space has well-defined characteristics, is a particular, has concrete relations to concrete things, and can explain concrete changes in things (such as change of shape after doubling). If space can deliver these and other sorts of theoretical goods for us, it does look like prejudice to insist on materiality and perceptibility. And clearly Leibniz shows us how well we can understand the characteristics of a space by appealing so ingeniously to a Euclidean character for it.

---

[9] See my "What Can Geometry Explain?" for illustration and discussion of this.

Could it ever be useful to include space in the ontology of some theory, or would we always be better off trying to avoid the postulate? The answer to this question is much less clear, though my own view is that space does some work in easily imagined theories, and does it elegantly enough to earn its keep. Relationist alternatives, by comparison, rob theoretical language of much of its assertive force and burden it with awkward conventions and indeterminacies. They impede a clear view of what the theory does, clutter it with nonfactual elements while offering no gain beyond the economy that they do not postulate space. Even then, I have been arguing, it is unclear how they really do avoid the postulate.

Leibniz's argument has loomed large because relationism needs to argue for the detachment thesis. The challenge to Clarke was no mere historical episode in which we now take only an antiquarian interest. It remains the most attractive relationist gambit. Relationism has no point without the detachment thesis. The argument of the challenge is fallacious, begging the question. On the other hand, if the discernibility principle is true, we have a strong argument from it to the meaningfulness of thing-space relations generally and so of space itself. It is easy to see how we might arrive at a posit of non-Euclidean space as our best explanation of what we observe.

GRAHAM NERLICH

University of Adelaide

AGENCY AND MORALITY*

AN "agent-centered restriction" (ACR), according to Samuel Scheffler,[1] is "a restriction which it is at least sometimes impermissible to violate in circumstances where a violation would serve to minimize overall violations of the very same restriction, and would have no other morally relevant consequences" (*ibid.*, p. 409). It would be wrong, for example, intentionally to kill one innocent person simply to minimize the number of innocent people intentionally killed. Similarly, breaking a promise, violating a right, or committing an injustice cannot be justified simply to minimize the number of identical broken promises, rights violations, or instances of injustice.

Because of this character, such restrictions harbor, for Scheffler, a "persistent air of paradox."[2] If a particular action is reprehensible, how can it (he suggests) be wrong to do it if, without other moral consequence, this will minimize the number of such actions. He notes that, although ACRs are central to much of our common morality, they apparently deny what seems to be the quite rational goal of minimizing what is morally objectionable.

Scheffler thinks this alleged conflict between our common (deontological) morality and a claim of reason creates a special burden for deontology; for, unlike ordinary consequentialism (e.g., utilitarianism), deontology rejects the principle that the moral good is to be maximized. We are not enjoined by its prescriptions to produce the most welfare, or right's protection, or virtue, or anything else (*ibid.*). And yet, how can it be wrong to torture one child (my example), if by so doing one simply prevents the identical torture of two children.[3]

---

* I would like to thank members of the Eastern Pennsylvania Philosophy Association and of the Philosophy Club of the University of St. Andrews, and especially my late colleague Seymour Schwimmer for help on earlier drafts of this paper.

[1] "Agent-Centered Restrictions, Rationality and the Virtues," *Mind*, XCIV (July 1984). Also see *The Rejection of Consequentialism* (New York: Oxford, 1982), pp. 80, 112. The nonmaximizing character of such restrictions is central. It is this property I am generally referring to in the paper when designating a moral point of view as "deontological," "nonconsequentialist," or "our common (or ordinary) morality." This is not meant to slight those personal commitments having to do with family, friends, etc. They certainly are a part of our "common morality." My aim, though, is to focus, as Scheffler does, on that powerful belief that it can be wrong to cause harm to prevent even greater harm.

[2] *The Rejection of Consequentialism*, p. 112.

[3] Scheffler apparently assumes about welfare or anything else conceived from a moral point of view that, if a particular action or consequence is bad, then, other things being equal, more instances are worse. Instance aggregation, however, though essentially unquestioned by Scheffler, may not be crucial to his challenge to

0022-362X/91/8804/190–212

Whatever makes torture objectionable, the torture of two children is worse than the torture of one. Moreover, if I torture a child, it is not, for that reason, worse than someone else torturing a child. Why should it matter (so the challenge goes) that one instance is caused by me if, without other moral consequence, that minimized instances of torture?

Scheffler, then, presents a striking internal challenge to deontology; 'internal' because of the suggestion that any moral perspective, including deontology, that identifies what is objectionable has reason to be an overridingly maximizing perspective. It would be as if to have a coherent conception of the decalogue one required an implicit eleventh commandment (or at least permission) that one could kill, steal, etc., in order at least to minimize the instances of killing, theft, etc. My aim in this paper is to answer the challenge by arguing for three claims: (I) there are three quite different conceptions of the importance of agency to moral obligation that should be, but often are not, kept distinct; (II) the particular sense of agency that Scheffler and others appear to believe is embodied in ACRs is irrelevant in accounting for their intuitive power; and (III) a patient- rather than agency-based defense of such restrictions is plausible and in accord with what, for Scheffler, is an important principle of practical reason. At the end of the paper, I make some further comments about agency in the light of this defense.

I. AGENCY AND MORAL OBLIGATION

Assuming the label 'agent-centered' (or 'agent-relative') is not simply a name for a nonmaximizing morality, we might wonder in what sense (if any) such a morality is connected to agency.[4] After all,

---

deontology. There are certainly cases where common morality forbids doing something to one person, although if you refrain, something worse is done to someone else. One cannot cut off the finger of one child to prevent the torture and death of another. In this paper, I accept for the sake of argument (with private reservations) that, if a particular action or state of affairs is morally objectionable, then more instances of the same are more objectionable. On this question of aggregation, see John Taurek, "Should the Numbers Count," *Philosophy and Public Affairs,* VI, 6 (Summer 1977): 293–316; Brook and Schwimmer, "On Adding the Good," *Social Theory and Practice,* VII, 3 (Fall 1981): 325–37; and Frances Myrna Kamm, "Equal Treatment and Equal Chances," *Philosophy and Public Affairs,* XIV, 2 (Spring 1985): 177–95.

[4] In some cases, Scheffler appears simply to stipulate that "agent-centered" is meant just to designate a nonmaximizing morality. "The agent relativity of these constraints and permissions consists in the fact that their application to individual agents is not based on or supported by any appeal to what would be best from an impersonal standpoint"—Scheffler, ed., *Consequentialism and Its Critics* (New York: Oxford, 1988), p. 5. But even if only a name, we would still want to know why 'agent-relative' was the appropriate one. And in a recent response to Jonathan Bennett on this question, Scheffler gives a more substantive (though I think mislead-

restrictions against harming some to aid (even more) others appear victim-centered, particularly if they are based (as I argue for later) on ascribing certain rights to people. In deciding whether or not to honor a right, I may have no concern that *I* rather than someone else might violate it. In his Tanner Lectures, Thomas Nagel[5] appears to conflate these considerations. He suggests rightly that deontological constraints begin with "claims of other persons not to be mistreated." Yet, against Nagel, it would seem precisely these claims which refer to agent-neutral values, since they concern states of affairs that give, in his words, "a reason for anyone to want or not to want" (e.g., not to have one's rights violated). Nagel says, however, that such claims give an agent the "agent-relative" reason "*not to maltreat others himself.*" Presumably, this reflexive use of the pronoun 'himself' is not simply redundant; what it suggests is that, for the agent, it is particularly important in honoring a claim that *he* not maltreat others rather than simply see to it that others are not maltreated (or, more broadly, not harmed).

This view about the significance of agency, which I call 'A1,' is given brief expression by Bernard Williams[6] when he says that "each of us is specially responsible for what he does rather than for what other people do" (*ibid.*, p. 117). Recently, Stephen Darwall[7] makes the comparable claim when discussing what he calls "agent-relative obligations":

> It is unquestionable that moral common sense suppose there to be *prima facie* obligations that are agent relative. The obligation not to

---

ing) answer to why deontological constraints have some significant relation to agency. I comment on this answer in the text. See Bennett, "Two Departures From Consequentialism" and Scheffler, "Deontology and the Agent, A Reply to Bennett," in *Ethics*, C, 1 (October 1989): 54–67 and 67–76, respectively. Jeremy Waldron, commenting on Robert Nozick's notion of "side constraints," suggests these constraints are "agent-relative" because each agent is taken to be concerned only with her own observance of the constraints"—"Rights in Conflict," *Ethics*, xcix, 3 (April 1989), p. 504. Literally, of course, this is false (as Waldron would undoubtedly admit) since we can certainly have obligations to prevent others from violating side constraints. Waldron's implicit suggestion, however, and the one I take up in the essay, is that calling a restriction "agent relative" means at bottom that one has a greater responsibility not to violate a constraint than to prevent someone else from violating the identical constraint (or preventing a comparable natural harm). I argue in the paper that this contrast is misleading because, although true, it is a derivative rather than a central point about deontology.

[5] "The Limits of Objectivity," reprinted in James P. Sterba, ed., *The Ethics of War and Nuclear Deterrence* (Belmont, CA: Wadsworth, 1985), pp. 15–23.

[6] "Utilitarianism and Integrity," in J. J. C. Smart and Williams, *Utilitarianism, for and against* (New York: Cambridge, 1973), pp. 76–117.

[7] "How Nowhere Can You Get and Do Ethics," *Ethics*, xcviii, 1 (October 1987): 133–57.

> betray a friend, for example, cannot be reduced to the neutral value of betrayals. If it could, it would be morally indifferent, other things being equal, whether one betrays a friend or does something that results, however indirectly, in a comparable betrayal in which *one is not directly involved at all* (my emphasis) (*ibid.*, p. 150).

Unlike Williams, Darwall is discussing the special duties of friendship, but his point is the same: although my betrayal of a friend is from an impersonal standpoint no worse than someone else's betrayal of a friend, I am more responsible not to cause that harm (betrayal) than to prevent others from causing an identical harm ("to prevent harms I am not involved in"). And, indeed, this claim that I am more responsible for my own actions than for the identical actions of others might, if correct, support the view that it is wrong to torture one child to prevent someone else from torturing two. A1 conceives moral obligation, then, in a personally indexed way, implying that the requirement that *one* not violate moral restrictions is more stringent than the demand simply to prevent violations.[8] A1 is, of course, too narrowly conceived if the contrast is only between my harming and my failure to prevent others from harming. Deontological restrictions, broadly conceived, forbid my harming to nullify threats initiated either by others or by nature. And, of course, natural threats are not violations.[9] Rightly conceived, then, A1 is the view that the duty not to cause harm is more stringent than the duty to prevent comparable harm.

The second sense of agency used to make a moral distinction, A2, is fundamental in the usual expression of deontological restrictions and is formulated (roughly) by the "doctrine of double effect": it is impermissible to cause a particular harm *in order* that (even more) good might occur, although it may be permissible in realizing that good knowingly but unintentionally to cause an identical harm. The doctrine has immense problems with respect both to its formulation and application, but it or some analogous distinction is historically central in expressing the nonmaximizing core of common morality.[10]

---

[8] It is striking how this conception of the importance of agency to deontology has gained currency in contemporary discussion. In a recent introductory text, we find the following: "An agent-centered pluralist morality could agree that deceit, coercion and injustice are bad. But it differs from consequentialism in giving each agent different aims; each of us should have the aim that *he* should not coerce or deceive other people, even if . . . [this] would make it the case that there was less coercion overall than there otherwise would have been. I shall follow tradition in describing agent-centered pluralist moral systems as *deontological* theories" (author's emphasis)—David McNaughton, *Moral Vision* (New York: Blackwell, 1988), p. 166. I mention later in a footnote an uncontentious sense in which deontology might be said to possibly give each agent different aims, while in another and obvious sense it gives each agent identical aims (see (fn. 21).

[9] Bennett (*op. cit.*) properly makes this point. Williams is, however, interested in the narrower contrast between responsibilities for my own actions and responsibilities for the free actions of others. This is the contrast raised in blackmail cases when I am asked to do evil to prevent someone else from freely doing more evil. I think these cases raise significant, distinct, and difficult questions, but Bennett rightly considers them too narrow to capture intuitively the restriction against harming some to prevent harm threatened by others or by nature.

The third sense of agency, A3, is a temporally indexed version of the doing-allowing distinction. It claims that *at a given moment* one is more responsible not to cause harm than to allow harm to occur. A1 and A3 are, unfortunately (as I argue below) not always distinguished in the usual puzzle cases about harming some to aid others.[11]

Clearly, A1 and A2 are quite different. Double effect draws a distinction within the compass of one's own agency; it permits balancing the good and bad effects of one's acts only if the foreseen but bad effects are unintended. There is perhaps something paradoxical here, but it is the paradox of distinguishing what is forbidden and permitted *for the same agent* by whether what is done is intended or only foreseen. Why should it be more permissible to discount the ends of another in foreseeing her harm as a consequence of one's action, than to discount the ends of another in merely using her for one's purpose? However we puzzle this out, A1 apparently ignores intention but, perhaps as paradoxically, suggests that, although a particular action is equally wrong whether done by *oneself* or *others,* one is more obligated not to do it than to prevent others from doing it.

---

[10] The Kantian distinction between using or not using someone as a mere means is a related distinction. This appears to be distinct from double effect, but see Onora O'Neill, "Kantian Approaches to some Famine Problems," in *Matters of Life and Death,* T. Regan, ed. (New York: Random House, 1980), pp. 285–94. O'Neill argues that to use persons as mere means is equivalent to intending to treat them in a way incompatible with their interests, therefore intending to harm them. For an interesting discussion of a set of distinctions with a family resemblance to double effect, see Christopher Boorse and Roy A. Sorensen, "Ducking Harm," this JOURNAL, LXXXV, 3 (March 1988): 115–35. See also the recent attempt by Warren Quinn to capture a more generic sense of mistreatment that is meant to include both strictly intending harm to a person and merely using a person which foreseeably harms her—"Actions, Intentions and Consequences," *Philosophy and Public Affairs,* XVIII, 4 (Fall 1989): 334–52. There is a great deal of literature, both religious and secular, on the doctrine of double effect, but it is not the main topic of this paper. Some other recent interesting discussions are in Charles Fried, *Right and Wrong* (Cambridge: Harvard, 1978); Nagel, *The View From Nowhere* (New York: Oxford, 1986), ch. IX; and Nancy Davis, "The Doctrine of Double Effect: Problems of Interpretation," *Pacific Philosophical Quarterly,* LXV (1984): 107–23. Most of this literature attempts some justification for a strong restriction on intentional harm. Very little of it offers a needed defense for the central claim of the doctrine that, in principle, merely foreseeable harm is permissible. Fried's work is an exception.

[11] Phillipa Foot expresses something like A3 when she says: "Nevertheless it is from the moral point of view a different thing to start up a new train of events in someone else's life and to refuse to intervene to stop one"; "Morality, Action and Outcome," in *Morality and Objectivity,* Ted Honderich, ed. (Boston: Routledge & Kegan Paul, 1985), p. 24. Since there is no temporal index in Foot's formulation, it is not clear she distinguishes the doing-allowing contrast from the responsibility for self versus responsibility for others contrast. For an interesting discussion of ambiguities in an earlier paper of Foot's, see Nancy Davis, "The Priority of Avoiding Harm," in Bonnie Steinbock, ed., *Killing and Letting Die* (Englewood Cliffs, NJ: Prentice Hall, 1980), pp. 173–215. Foot's earlier piece is "Abortion and the Doctrine of Double Effect," repr. in J. Rachels, ed., *Moral Problems* (New York: Harper & Row, 1975), pp. 59–71.

A1 and A2 can be ambiguously lumped together. Take, for example, the following illustration of Nagel's: "And it seems you shouldn't twist the arm of a small child *to get* its grandmother to do something, even if the thing is quite important—important enough so that it would not be reasonable to forgo a comparable benefit in order to prevent *someone else* from twisting a child's arm" (my emphasis).[12] Both senses of agency are suggested: intention (twisting the arm to get a result) and the alleged special responsibility for one's own actions as opposed to the actions of others (not twisting the arm yourself as opposed to preventing someone else from twisting the arm). Yet Nagel ultimately opts for the distinction between intention and mere foresight as embodying what he calls "agent-relative values," values that express, as he says, the "particular relation of the agent to the outcome" (*ibid.,* p. 17). He contends that the simply foreseen consequences of one's actions [are] "not in the *relevant sense* what one *does,* and [do] not come under a deontological constraint" (*ibid.,* p. 19; my emphasis).

This claim that the foreseen consequences of my actions are not "in the relevant [moral] sense" what I do assumes without argument (here at least), however, that such consequences should not be subject to constraints as strict as those on consequences intended. That point would not automatically follow even if it could be made out that instrumentally harming was intrinsically worse than consequentially harming. Double effect is at bottom a doctrine about what is permissible. Therefore, to make his claim, it is not sufficient for Nagel to remark on some special reprehensibleness in "aiming" at evil (e.g., the child's pain) as an instrument for good.[13] This point, though suggestive, tells us little why, if unintended, we can permissibly cause an identical harm.

---

[12] "Agent Relative Morality," in *The Ethics of War and Nuclear Deterrence,* p. 19.

[13] *The View From Nowhere,* pp. 176–80. Nagel's point is that, in intending to cause the child pain, one "tracks" the pain, twisting the arm harder to increase it if that is needed. But we can imagine something quite similar about unintended but foreseeable harm. For example, suppose that, if we warn civilians who might be killed as a collateral but unintended effect of bombing a military installation, the installation would be moved. So we cannot warn the civilians. We might, in fact, do everything to ensure they stay. We in effect track the civilians and ensure that they die, though we do not intend their deaths. In principle (subject to proportionality constraints), double effect permits this.

In any event, by labeling deontological restrictions 'agent-centered', Scheffler is not merely referring to double effect's distinction between intention and foresight. His challenge is to the rationality of forbidding one violation of a restriction though that serves, without other consequence, to minimize identical violations. Double effect, however, marks the primary moral difference between types of things one does rather than between consequences caused and consequences not prevented. Scheffler's examples and discussion strongly suggest that the sense of agency he thinks embodied in ACRs is A1.[14] Commenting on the "ordinary view" of the contrast between negative and positive duties, he says: "But if *you* would have to murder one innocent person to prevent the murder of five by someone else then you must not do so."[15] And in the following passage, he gives "a schematic example" of the alleged paradoxical character of ACRs. "Suppose that if agent A1 fails to violate a restriction R by harming some undeserving person P1, *then five other agents A2 . . . A6*, will each violate restriction R by identically harming five other persons, P2 . . . P6, who are just as undeserving as P1, and whom it would be just as undesirable from an impersonal standpoint to have harmed" (*ibid.*, p. 84; my emphasis). This suggests that Scheffler, like Williams, Darwall, and Nagel, thinks ACRs involve a personally indexed view of the importance of agency, i.e., that the agent conceives her special obligation that *she* not do wrong (as opposed to just seeing that wrong not be done), where the contrast emphasized is between obligations for one's own actions and obligations for the comparable actions of others.[16] Scheffler makes the same point in a recent exchange with Jonathan Bennett. Discussing the "doing-allowing" distinction, he correctly notes that that distinction (or Bennett's "making-allowing" distinction) cannot be fundamental to deontological restrictions. If it were, consequentialism gains support, for then it seems plausible to permit doing *X* (or making *X* happen) if only to reduce the number of doings (or makings) of *X*. Scheffler puts it this way: "the significance of that distinction [doing-allowing] is agent-relative in character; what matters are the agent's doings, not the total number of doings overall."[17]

---

[14] Scheffler, of course, is not defending A1. As a matter of fact, he is quite skeptical about its worth. I note, as well, that double effect's strong proscription against intentionally causing harm would be—using Scheffler's terminology—a species of agent-centered restriction, since one is forbidden to harm intentionally even to prevent more intentional harm.

[15] *The Rejection of Consequentialism*, p. 24. The italicized 'you' is interesting. There is a difference between: (1) *you* should not kill; and (2) you should not *kill*. The first appears to make moral restrictions relative to the agent. The second more plausibly expresses the general prohibition against killing which distributively applies to any agent.

[16] This is the way Christine Korsgaard interprets Scheffler in "Personal Identity and the Unity of Agency: A Kantian Response to Parfit," *Philosophy and Public Affairs*, xviii, 2 (Spring 1989), p. 102. An agent-centered morality, she writes, requires that "you must be in a special way concerned with what *you* do."

Presumably, Scheffler means that what matters particularly to the agent is the agent's doings as opposed to the doings of others.[18] And it is this conception of differential responsibility which I am labeling 'A1'. Yet, significant as A1 appears to be in accounting for our ordinary (deontological) moral intuitions, I argue below that it is derivative rather than central to the force of those intuitions and therefore cannot be central to any justification for deontological constraints.

II. IS A1 ESSENTIAL TO DEONTOLOGY?

Consider the following case. You are at the zoo with two children who are making a scene. Becoming angry, you toss them into the lion's den. Horrified, you come to your senses and notice that they can only be saved if you toss a third child (who just toddled along) into the back of the den. The beast would be distracted and you could quickly leap in and save the first two. What should you do? Whatever you do, you have initiated a chain of events that results in a child's death. If you do nothing, you have intentionally killed the first two children. If you sacrifice the third child, you kill it in using it as a mere means to save the others. Or consider this variation of Darwall's example about betrayal. Captured by the enemy and under some pressure, I reveal the hiding place of two comrades. They will be killed if discovered. I could prevent this by revealing the whereabouts of a third comrade, perhaps of more importance to the enemy. Whether or not I reveal her whereabouts, I have intentionally betrayed someone.[19] And equally important, if I fail to betray the third comrade *I still remain directly involved* in the betrayal of the first two.

---

[17] "Deontology and the Agent," *Ethics*, C, 1 (October 1989), p. 73.

[18] There is, of course, a minimum and uncontentious sense in which deontology may give each agent different aims. The state of affairs described as 'Smith not intentionally killing an innocent person' can refer to an aim of Smiths but not necessarily to an aim of Jones. The personal index is read into the content of the aim. Ordinarily, however, the personal index is read as an operator—each person has the identical aim not to kill an innocent person. Having the index in the content of the aim suggests, misleadingly, that a belief about differential responsibility—that one has more responsibility for one's own actions than the comparable actions of others—is central to deontological restrictions. But that it is central (as opposed to being derivative) certainly does not follow from reading the personal index into the aim. That reading leaves it open what Jones's responsibility is with respect to the aim described as 'Smith not kill an innocent person'.

[19] Of course, in this example more than one agent is involved. The important identity of structure in the examples, however, is that the person who initiated the

Yet it seems clearly wrong to kill the third child or betray the third friend; as wrong, in fact, as in any of the standard cases where you might harm one person to aid others. In terms of what is important, these are, in fact, standard cases. It is perhaps just an artifact of the usual examples that agency (in the sense of A1) appears to be important. Perhaps we are more likely to face harming some to nullify threats initiated by others (or by nature) rather than by ourselves. But that difference in the origin of the threat is irrelevant to the intuitive force of the prohibition against harming some to aid others.

Of course, A3 might justify not killing the third child or betraying the third friend. One can claim the agent has a greater responsibility *at a given moment* not to violate a constraint than *at that moment* to prevent a violation. The temporal index here, however, is crucial. Without it, the doing-allowing distinction cannot provide a rationale for not killing the third child. Whether or not I kill that child, I have done something, killed children, rather than merely allowed them to die. Admittedly, Scheffler's definition of an ACR on occasion appears compatible with either A1 or A3.[20] Yet the latter, I think, fails to connect agency in any significant way with agent centeredness; that is, it does not connect moral restrictions with one's agency interpreted as a special concern with one's own doings overall. After all, what counts as what I do extends over more of my lifetime than to what I do at a particular moment. It is true that the temporally indexed version of the doing-allowing distinction would, in accord with common morality, forbid my murdering one child though that *allows* it to be the case that I have murdered two others. But A3 is far from being "agent-centered" in the sense of giving me a special obligation that *I* not violate deontological constraints. If a restriction were truly agent-centered in this sense, each agent would have the plausible goal of minimizing her own violations as opposed to total violations.[21] In the case of the children, however, that is precisely what deontology forbids. It is wrong to murder the third child to nullify my previous attempt to murder two. Single-agent examples (where the person who can nullify a threat, originated the threat)

---

threat is the same person who might nullify it by harming someone else. Under A1, in my version of Darwall's example, I would have the same responsibility for the fate of any comrade since after the first betrayal I am responsible for a death whatever I do.

[20] For example, the passage quoted at the beginning of this paper. And, as I mentioned (fn. 4), he suggests in a recent piece that 'agent-relative' simply designates a nonmaximizing morality. But this raises again the question why, in a substantive sense, nonconsequentialism is an agent-relative point of view.

[21] I discuss later in the paper the view that intention rather than consequence determines wrongness. In not doing that here I think I am being fair to the spirit of Scheffler's notion of constraint violations.

allows, in fact, a criticism of Scheffler similar to his criticism of Bennett. Scheffler's point against Bennett was that, if the doing-allowing distinction were fundamental, one reasonably should be permitted to do *X* to minimize the total number of doings of *X*. But in the same way, if an agent were overridingly concerned with her own doings, a limited consequentialism equally gains a foothold. The agent plausibly should be permitted to do *X* to minimize her own doings of *X*. But, again, deontology forbids this.

We might combine both the temporal and the personal indexing and say the agent, at a given moment, has a special obligation that she (at that moment) not cause a violation. But the two forms of indexing work at cross purposes. My personal responsibility again is substantially coextensive with my identity over time as an agent. At any given moment *t*, I am as responsible for the consequences of actions I initiated prior to *t* as for those I begin at *t*.[22] The temporally indexed version of the doing-allowing distinction, however, discounts the general significance of my agency since, at a given moment, it makes me more responsible not to violate a moral restriction then to see to it (at that moment) that *I* do not end up violating more restrictions. I do think (and defend later) that A3 remains significant for deontology, but it is of derivative rather than primary importance.

In any event, the personally indexed sense of agency, A1, which appears to be what Scheffler means by "agent centeredness," cannot account for the conviction over the full spectrum of examples that it is wrong to cause serious harm to some to nullify threats of more harm to others; for the compelling power of this conviction remains when one considers cases where one is the only agent. In these cases, no contrast is made between my obligations not to harm and my obligations to prevent others (or nature) from harming. Yet it is, for ordinary morality, as wrong for me to torture one person to nullify either my own or another's previous threat to torture two.[23]

What remains distinctive about deontology, then, is not some personally indexed view of moral responsibility, but simply its nonmaximizing character. And, indeed, this character is perhaps even more paradoxical (as Scheffler[24] notes) if A1 is rejected. For if it is morally irrelevant who (or what) brings about a consequence, it might seem that only the consequence matters. Of course, the irrelevancy of A1 does not rule out an agency-based (if not agent-centered) defense of common morality using either a temporally indexed version of double effect or of the doing-allowing distinction.[25] Another possibility, however, is a patient-based defense of deontology. This is the direction I take in the last section of the paper.

---

[22] This should be amended somewhat. If an agent radically changes personality, we might say she is no longer morally responsible (as opposed to physically responsible) for prior actions. But in any case my moral responsibility, though perhaps not coextensive with my physical agency, includes more than the consequences of actions I initiate in the present.

[23] Although the stuff perhaps of science fiction, we could imagine that Smith has started a device that, if not stopped, will torture two people. Only the scream of a third person, induced by a certain amount of pain, has the required pitch to destroy the device. I owe the idea for this type of example to a discussion of double effect with Nozick.

I would note first that, in responding to some criticism by Phillipa Foot, Scheffler denies that his challenge requires some conception of an "overall best or worst state of affairs," as, for example, in utilitarianism. Rather, he contends that consequentialism embodies a deep and quite general principle of rationality which he calls "maximizing rationality" (MR):

> The core of this conception of rationality is the idea that if one accepts the desirability of a certain goal being achieved, and if one had the choice between two options, one of which is certain to accomplish the goal better than the other, then it is, *ceteris paribus*, rational to choose the former over the latter.[26]

Agent-centered restrictions appear to violate this principle because they claim certain states of affairs are bad (e.g., a child being tortured), but (*ceteris paribus*) forbid acting so as to minimize the number of such states. "How," Scheffler asks, "can it be rational to forbid the performance of a morally objectionable action that will have the effect of minimizing the total number of comparably objectionable actions that are performed and have no other morally relevant consequences?" (*ibid.*, p. 413).

Scheffler allows that deontological restrictions could be substantially compatible with MR but only if the objectionableness of an action's consequence does not exhaustively account for the restriction (*ibid.*, p. 418).[27] It is interesting that the additional element he thinks might be required for compatibility appears to be something like A1. He suggests, for example, that, if for each agent, the fact that *she* was violating a restriction was an element in the objectionableness of the violation, then each agent presumably might have the coherent agent-relative goal that she not violate a restriction (as opposed to the goal of minimizing the total number of violations). This is not a view which he thinks credible, since he thinks it conflicts with the impersonal character of moral aims, but a view, he believes, which would at least make deontological restrictions consistent with MR: "views that include agent-centered restrictions need not come into conflict with maximizing rationality when they tell us to further the agent relative goal of not violating the restrictions *ourselves* at the expense of the non-relative goal of minimizing violations of the restrictions" (*ibid.*, p. 417; my emphasis).[28]

---

[24] "Agent-Centered Restrictions, Rationality and the Virtues," p. 414.

[25] In order to work in single-agent examples, double effect would also have to be temporally indexed, i.e., what at a given moment I intend or simply foresee. In the children and the lion example, I violate the nonindexed version of double effect whether or not I kill the third child.

[26] "Agent-Centered Restrictions, Rationality and the Virtues," p. 414. Foot's discussion is in "Utilitarianism and the Virtues" and "Morality, Action and Outcome," in *Morality and Objectivity*.

[27] I say 'substantially' to put aside the truism that, if an agent's goal at a moment is not to violate moral restrictions, then MR is automatically satisfied by her acting in accord with deontology's constraints. But this response masks the significance of Scheffler's challenge that, once the agent's goal is given a more substantial description, e.g., 'don't use anyone merely as a means', our attention is directed to a particular consequence (being used in a certain way). There then seems some intuitive pressure to minimize such consequences.

Against Scheffler, I have argued that A1 (though perhaps in accord with MR) is not compatible with restrictions against harming some to aid others. An agent taking seriously A1's view of moral responsibility could reasonably choose to minimize her own total violations (as opposed to all violations), and thus to do what ordinary morality forbids, kill one person to nullify her prior attempt to kill two.

Scheffler, then, finds the constraints of ordinary morality paradoxical, because he apparently thinks A1 is required for them to be consistent with MR, but believes A1 to be unjustified (*ibid.*, p. 415).[29] I do agree that A1 appears to conflict with the impersonal character of moral requirements. It is no element (by itself) in the wrongness of an act that I do it rather than someone else. But, of perhaps more interest, A1 is not even compatible with deontological constraints since it appears to permit violating a constraint to reduce one's total violations of the same constraint. In fact, it is this failure of A1 to support nonconsequentialism which suggests the latter needs a patient- rather than agency-based defense.

III. A PATIENT-BASED DEFENSE OF DEONTOLOGY

To answer Scheffler's challenge, we must locate some property P of a patient for which the following propositions are true: (1) the existence of P explains why a state of affairs characterized by P is morally objectionable; (2) if a subject S ends up bringing about two instances of *P*, she has done (in terms of the consequences of her act)[30] something worse than another subject *S** who brings about one instance of *P* (*S* can be identical to *S**); and (3) the impermissibility of causing one instance of *P* simply to prevent more instances of *P* is in accord with practical reason. Take the case, again, where I consider killing one child by using its body to distract a lion from two I previously endangered. Regardless of who endangered the first two, it seems quite wrong to kill the third child. What might account for the force of this intuition? The case can be tightened up by making it "one on one." Assume I tossed one child into the den, but could save him by distracting the lion with a second (who then would be killed). Whether I sacrifice the second child or not, I have unjustly killed a child. Having no concern now about the numbers, one might seek some property possessed by the second child to defend a principled refusal to kill it.

---

[28] Scheffler argues here, persuasively I think, that this personally indexed sense of agency is not compatible with taking "the moral point of view." This appears to be the basis for his view that there is some "*prima facie* irrationality" in deontology; *Rejection of Consequentialism*, p. 83.

[29] The point he makes here is that the moral (as opposed to the egoist) point of view must take a stand about the objectionableness of states of affairs (say a child being tortured) independent of agency.

The obviously distinct property of the second child is that it has not at that moment been significantly threatened.[31] I mean by this at least (and somewhat roughly) that no causal chain has been initiated which, if uninterrupted, will imminently lead to its harm.[32] The second child might well have a disease that will inevitably kill him in twenty years. But this does not alter his status as significantly unthreatened when I consider using his body to distract the lion. Another way to put this is to say that, even with the disease, the second child is significantly benefited by my not injuring or killing him to save the first.[33] As an imaginative device, we might think of a moral physics where the life lines of people pursuing their interests but not threatening others are analogous to force-free or inertial motions in mechanics. As in a natural physics no explanation is required for inertial motion, so here no justification is required for individuals to pursue their nonthreatening interests. And just as in mechanics deviation from inertial motion requires an explanation, distorting the path of an individual by threat requires justification. There is a moral shield, possessed by each person, a shield against threats, i.e., actions that change the world so as to bring an individual into harm's way. The fundamental notion is that not being significantly threatened at a given moment is a basic and morally important property.[34] The child I might kill to save the first possesses at that moment the property of not being imminently threatened with death, a property lacked (then) by the first child.

---

[30] It is important to emphasize that (2) only refers to consequences of an action. And by 'consequence' here I mean anything that could be the referent for a consequence description, e.g., 'harming a person', or 'violating a person's right'. There is, of course, a dimension of "wrongdoing," that allows grading actions in terms of which is worse than another in terms of *mens rea* rather than consequences. I discuss this later in the paper as a possible response to Scheffler. This note was prompted by some comments of an editor of this JOURNAL.

[31] Except for the Pickwickian sense in which the second child is threatened because I might decide consequentialism is correct.

[32] A difficulty here is how to consider sincere threats from, say, terrorists. We might not want to consider such threats as themselves the initiation of causal chains that, if uninterrupted, will lead to harm. In the case I am considering, however, the threat is already embodied in an attempt at harm.

[33] What of cases where, if you did not kill a person to save others, that person and the others would be immediately killed by someone else? That person is not benefited by your refusal to kill him. Williams pursues this kind of case, suggesting that the moral importance we place on our agency accounts for the belief that it may still be wrong to kill the one. Williams, however, does not make much of the aspect of overdetermination, i.e., that the one you might kill would die in any event. His point is to contrast one's responsibility for one's own actions with responsibility for the free actions of others. Thus, it is not clear to me how Williams would think of examples where either (a) the one would not die if you did not kill him, or (b) if you did not kill one, all would be killed by natural causes (or a process already initiated by another); see "Utilitarianism and Integrity."

It is important in this regard to remember that Scheffler's challenge to deontology contains a *ceteris paribus* clause: he suggests that, if a state of affairs, *S*, is objectionable and other morally relevant considerations are equal, it seems reasonable to minimize the instances of *S*. Certainly, the clause is satisfied if my choice is simply to prevent one murder or two. Then (allowing adding the value of lives) I should prevent two. On the other hand, if I have to murder one child to prevent my having murdered another, then the *ceteris paribus* clause is violated. But not—and this I think quite important—because of the moral significance of my agency. What is significant is not that I (as opposed to someone else, or some natural cause) would cause the death but rather the protection due the child who at that moment is not threatened.

The morally significant feature of not being threatened at a moment demands some special protection, a protection embodied in the right not to be harmed. Expressing the right as one simply against threats (as opposed to the consequences of threats) appears too weak and is in fact unnecessary, since, except for special (though interesting) cases, harming a person entails having altered the world in a way that significantly threatens her with harm.[35] Thus, the right not to be harmed makes concrete the special protection against threats of harm.

---

[34] Obviously, there is some idealization here. A person might be unthreatened at a moment but deserve harm for some past action; or the person might simultaneously be unthreatened and wrongly threatening others; or he might be innocently threatening others. I exclude such cases as we might ignore various disturbances in an idealized natural physics. This is not to deny their importance.

[35] The exceptions I have in mind are cases of overdetermination, e.g., drowning a person who has fallen out of a boat, or stealing money from a person who is fleeing a thief. These seem to be examples where harming a person does not entail originating the threat (at least in a generic sense) of harm. But in the standard puzzle cases where harming some is required to aid (more) others, the act of harming originates in some significant sense the threat of harm.

But however we think of the content of rights, what is important is that a protection expressed as a right requires a principle of "non-trade-off," i.e., one cannot violate a right simply to minimize the number of identical right's violations. Although (some if not all) rights can be overridden, minimizing the number of identical violations is not, in itself, a sufficient reason to violate a right.[36] Although the "non-trade-off" property is an accepted feature of rights, it is not just being stipulated here. In the case of the children, the non-trade-off feature expresses the special protection due the then-unthreatened child. If no special protection were due that child, it would be immaterial in the "one on one" case, other moral considerations being equal, whether I killed the second child to save the first, or did not kill the second child and thereby killed the first (I could flip a coin). Either way, I minimize violations of the right not to be killed. Contrapositively, if some special protection is due the unthreatened child, then it is wrong for me to kill the second child to prevent my killing the first.

Of course, consequentialists might argue that moral considerations are unlikely to be equal in the "one on one" case. Not only is there no net gain in lives by killing the second child to prevent one's killing the first, but there is some cost in the effort to do that. Thus, by a utilitarian calculus, it might come out wrong to kill the second child without assigning it any special moral status *vis à vis* the first. But, after all, the cost to oneself of the effort to kill the second child might be balanced either by some slight preference for the first child or by something else. We can imagine examples, then, where the remaining relevant factors, aside from the question of protection, are only whether I kill the second child or allow it to be the case that I have killed the first.

---

[36] I note a comparable conception of rights in a recent article by Phillip Petit. One feature Petit stipulates is a "non-tradable across persons" feature of rights, explained as "A's right to X is not overridden simply because his [the agent's] violating A's right would minimize the violation of people generally of the right to X on the part of people generally"; see "The Consequentialist Can Recognize Rights," *The Philosophical Quarterly*, XXXVII, 147 (April 1987), p. 46. My conception of a "non-trade-off" feature of rights has some similarity to H. M. Malm's principle that "It is morally impermissible to substitute one person's death for another's (change who lives and who dies) without a good reason for doing so"; "Killing, Letting Die and Simple Conflicts," *Philosophy and Public Affairs*, XVIII, 3 (Summer 1989), p. 248. Malm's principle is weaker than Petit's but at least requires that a good reason for killing cannot be simply that the killing resulted in a situation (impersonally judged) no worse than not killing.

One might also argue that, even without claiming special protection for the second child, there is no reason (particularly if other considerations are balanced) to kill that child to save the first. And, it is said, one needs a reason to act. This argument claims that my somewhat technical sense of 'minimize'—that, whether I kill the second child or not, I have the same (minimum) number of violations—is tricky perhaps, but not to the point. On the other hand (the objection continues), if my violation actually minimized violations in the sense of reducing them, one would in fact have a reason to act. The argument has a point. But it seems to me that saying an action is unreasonable is not equivalent to saying it is forbidden. Consequently, if (when other considerations balance) minimizing right's violations were sufficient reason to violate a right, it still remains permitted, if perhaps unreasonable, to kill the second child to save the first. Moreover, for consequentialists, the mere fact that one brings about equivalent harm by doing *A* or by not doing *A* is not a moral reason for refraining from doing *A*. And, for nonconsequentialists, what makes killing the second child forbidden is not its unreasonableness, but rather the special status of the unthreatened child, a status embodied in the right not to be killed.[37]

It cannot then be correct to say, as David Brink[38] does (explicating Scheffler), that "if a violation of a right is a *bad thing* then it is rational to want to minimize the violation of rights even if we must violate one person's rights in order to do this" (*ibid.*, p. 436; my emphasis). There is an illusion of an intuition here; an illusion that is simply an artifact of the use of evaluative language. The "conjuring trick" has been accomplished if the requirement of reason to minimize is gained by just noting that the violation of a right is a "bad thing." The general concept "bad thing" just will not do that work. The violation of a right is in truth a bad thing, and bad for that reason (that it violates a right). And if Smith violates three particular rights, he perhaps does (to that extent) something worse than Jones who violates only one. But, given the protection of a right, it can remain forbidden to violate one right simply to minimize right's violations.[39]

---

[37] Of course, the first child has exactly the same right. But given the non-trade-off feature of rights, the first child's right is moot in situations where I have to kill another to preserve that right.

[38] "Utilitarian Morality and the Personal Point of View," this JOURNAL, LXXXIII, 8 (August 1986): 417–39.

[39] A more sympathetic interpretation is that Brink is merely claiming that, if the violation of a right is a bad thing, then *ceteris paribus* it is rational to want to minimize right's violations. But this might just be a truism and certainly removes the bite of his point. It leaves open the possibility (one I argue for in the text) that

But suppose we add children in the example above. By killing one child, I could save two, or three, or, in general, N children ($N > 1$) that I have previously threatened. Would that make a difference? This is a difficult but a different question. Some rights can be overridden in some circumstances. But not, as the "one on one" case attempted to show, in order merely to minimize right's violations. Questions remain, of course, about the extent of the protection to those not threatened. I am inclined to think the number of cases of identical harm one would prevent by harming is irrelevant. Others might think that, if one's murder prevented ten identical murders or accidental deaths (but no less), it would be justified. We might say, in general (and admittedly vaguely), violations are permitted if they bring about significantly more good. And although there could be general criteria here, there may be no precise boundary that establishes significance as there is no precise boundary between short and tall.[40]

The special protection due X does entail that I cannot violate X's right to life just to nullify my prior and unjust attempt to kill Y. Admittedly, it does not necessarily follow that it would be wrong to kill X to nullify my prior threat to kill Y and Z. But once we allow some special protection at a moment to those who are significantly unthreatened, there is no Schefflerian "paradox" about an imperative that forbids killing one child to save one, two, or more children from either murder or accidental death. That paradox depended on concluding that, if agency was irrelevant to obligation, then one's causing a violation should be in itself morally irrelevant. I have agreed that agency (at least A1) is irrelevant to obligation but that the protection due the presently unthreatened person is what makes one's causing harm morally relevant. In rejecting the paradox, the door is now open to claiming that a violation must produce significant good rather than simply that it minimize identical violations. Of course, achieving some significant good at the cost of a violation may turn out (in a particular case) also to minimize violations. But it is not constitutive of being a significant good that that occur. The violation

---

violating a right to minimize violations of the same right violates the *ceteris paribus* condition. Brink's use of the category "bad thing" is misleading in another way. If one focuses on welfare, for example, maximizing doctrines like utilitarianism do have some attraction. On the other hand, that attraction is significantly diminished when focusing on something as general as moral disvalue, a category that could include right's violations, failed attempts to harm, unrequited lust, or simply slothfulness. Put simply, there is no moral demand to minimize immorality.

[40] For an interesting discussion of how this notion of "significantly more good" might be worked out see Kamm, "The Choice Between People: Common Sense Morality and Doctors," *Bioethics* I, 3 (1987): 256–71.

of a right, then, is the property, *P*, which can refer to the consequence of an action and which answers Scheffler's challenge to find some property which both explains, without reference to agency, why a state of affairs is morally objectionable and yet which entails that it can be forbidden to cause one state simply in order to minimize the number of identical states.

An interesting objection to the above argument is that, by making deontology's nonmaximizing character formal (with the non-trade-off principle), we just make explicit rather than deflate its alleged paradoxical nature. The paradox, some might argue, remains the conflict between this formalism and a more general and better grounded principle of practical reason, i.e., MR.[41] For example, we have admitted that an agent by not committing one violation of a constraint may end up doing something worse, violating two constraints.[42] This appears at odds with the principle of maximizing rationality.

The point would be cogent if, in fact, MR (as Scheffler defines it) requires the interpretation he gives it. Scheffler's definition of MR, however, which ultimately is a principle of instrumental rationality, allows a variety of interpretations from the admittedly vacuous "an agent best achieves the goal of obeying the rules of morality by obeying the rules of morality," to the interpretation that he appears to favor, that, if a state of affairs is morally objectionable and it is morally irrelevant who causes it, then it seems reasonable to cause one state at least to minimize the total number. And there are at least two other interpretations of MR (I can think of) which are not vacuous.

For example, in a moral tradition where what makes an action objectionable is intent (or foresight, or negligence) rather than actual consequences, one can, in fact, minimize what is objectionable by not harming some to save a greater number from one's prior attempts at harm. Those prior attempts constitute wrongs and cannot be undone.[43] If an agent was required to minimize wrongdoing in this sense, then MR is in accord with ordinary deontological constraints. In fact, some commentators on an earlier draft of this paper[44] think this is the simplest way to handle Scheffler's challenge. They contend that the agent does maximize wrongdoing by murdering one child to nullify her prior threat to murder two. The prior threat constituted two violations of a constraint (regardless of whether a death ensued); the murder is a third violation. This view strictly thinks of wrongdoing, including violations of rights, in terms of *mens rea* rather than unjust harm. From this perspective, it is necessarily true that an agent who at any given moment obeys a constraint not to harm some to nullify her own or other's prior threat of harm minimizes wrongdoing overall.

---

[41] This point was made to me by Richard Montgomery and John Whelan.
[42] This was a point raised by an editor of this JOURNAL.
[43] I owe the suggestions along these lines as well as other helpful comments to Anthony Ellis.
[44] Both Ellis and David Carr suggested this.

Divorcing intention (negligence, recklessness, etc.) from consequences in the above fashion, and associating wrongdoing only with the former, makes it a matter of course, at least in the trade-off cases discussed, that an agent obeying the constraints of deontology minimizes wrongdoing. So some maximizing demand is met. I have some misgivings, however, about using this argument to show deontology is compatible with MR as Scheffler defines it. MR (for Scheffler) requires a particular goal, but it is not the goal of a deontological agent to minimize her own or anyone else's wrongdoing. Rather, it is simply that, if *mens rea* defines wrongdoing, then an agent obeying a restriction against harming some to aid others will in fact minimize wrongdoing.

Even if we conceived right's violations as requiring actual harm rather than merely attempts at harm, there is, I think, a fourth interpretation of MR (as Scheffler defines it) independent of agency and in accord with ordinary deontological restrictions. One can have the goal, at a given moment, of not violating a right without believing it is of special significance that oneself as opposed to others not violate the right. And the agent can quite rationally believe, given the special protection due the unthreatened, that violations are forbidden just to minimize instances of identical violations. Scheffler apparently believes that, if an agent prefers no right's violations to one, she should prefer less violations to more, unless what is significant for her is that she not violate a right.[45] But this ignores the possibility that resistance to the goal of minimizing total violations (the goal he believes required by MR) follows from the nature of rights rather than from special concern with one's agency.

It seems to me, then, that MR—choosing the best means for a desired goal—does not entail, without begging the question in favor of consequentialism, any demand to minimize my own or others' right's violations. The goal in the conception I have presented is temporally indexed to respect at any given moment the right of one who is innocent and unthreatened. An agent best meets the goal by respecting the right at that moment, though as a result she may end up having done something worse overall, for example, having violated the rights of two people. We can even say that the agent does something worse at the moment by violating the third person's right than by not violating it. In this sense an agent, by not violating a right at a given moment, does minimize her own wrongdoing at that moment. But, again, this is not because of some special moral significance to her agency, but rather because of a temporally indexed goal that reflects the right of a patient. And it is precisely that index which allows us to resist the notion that, implicit in a commitment not to violate a right, is an agent's commitment to reduce her own or others' right's violations overall. Thus, an agent can both *do* something worse at a moment by violating a right than by not violating it and end up *having done* something worse by not violating the right. That is, the agent, by not violating one right at time *t*, can end up having violated two rights. But at *t* she does not violate two rights, but allows that to occur. There would only be a paradox here if (as Scheffler appears to believe) implicit in a conception of wrongdoing is either an imperative of practical reason to minimize wrongdoing overall, or (as he thinks true, but irrational, about deontology) a demand on an agent to minimize her own wrongdoing. I have suggested a third alternative, a temporally indexed commitment that reflects the right of an unthreatened patient.

---

[45] See, again, "Agent-Centered Restrictions, Rationality, and the Virtues," p. 415.

FURTHER COMMENTS ABOUT AGENCY, DOING VERSUS ALLOWING, AND DOUBLE EFFECT

As for A3, the temporally indexed version of the doing-allowing distinction, it retains only derivative significance in a patient-based defense of deontology.[46] Violating constraints against harming some to aid others requires intentionally or knowingly (or negligently) at a given moment altering the world to threaten harm to one then significantly unthreatened. Therefore, violating a constraint at a moment is *to do* something rather than (at that moment) *merely allowing* it to occur. Of course, even merely allowing something, *X*, to occur may invariably entail that one does something else, *Y*. In allowing the general to drown, I have done something, altered the course of the war, although I have not necessarily killed the general. But the contrast I want is between doing *X* (by commission or omission) at a moment and merely allowing at the moment more instances of *X* to occur. That contrast (between doing and allowing at a moment) is needed for the puzzle cases about harming some to prevent the same harm to others even to get started, and is based on whether before one's action (or inaction) another is significantly threatened with *X* happening to him.[47]

---

[46] That it must remain significant was pointed out to me by Ellis.

The temporal indexing in A3 is required to support formally the strong belief that it can be wrong at a moment for an agent to cause harm to nullify her own prior attempts at more harm. Without the indexing, the agent (in single-agent cases) is faced simply with a (temporally unrestricted) choice of how many people ultimately to harm. But the temporal indexing is neither fundamental nor an arbitrary device to "save the (moral) phenomena." It is based on the special protection due those who (at the moment) are both innocent and unthreatened.

One objection to the above comments is that one gains little by calling my conception a patient-based deontology, since in the cases considered there is a "reciprocal entailment" between the duties of agents expressed by A3 and the rights of patients.[48] Why not allow A3, then, to be fundamental, and deontology to be agent- rather than patient-based. The alleged "special protection" due the innocent merely reflects, on this account, the fundamental duties of agents at any given moment not to bring the innocent under threat.

The answer, I believe, is that, though there is a reciprocal entailment of rights and duties in the cases I have discussed, the order of explanation is asymmetrical. It makes sense to explain the temporally indexed duty in terms of the special protection due someone who is unthreatened, and not the other way around. That right is supervenient on the status at a moment of an innocent person's pursuing her unthreatening interests. This is the default position. It is the moral significance of that status that accounts for the fact that the agent's duty must be temporally indexed. Otherwise, we would wonder why a person especially concerned with her own agency would not wish to minimize her own wrongdoing overall (i.e., in a temporally unrestricted way). The temporal index reflects the moral status of the unthreatened person whose life the agent might (but should not) trade to save others she has previously threatened. The order of explanation is from this status to the duties of agents. It is in this sense that I consider deontology patient-based. The distinction I wish to draw, again, is between this view of deontology and one which Scheffler seems to espouse, which conceives an agent concerned ultimately with her own doings as opposed to overall consequences.

---

[47] Omissions as well as commissions can certainly count as "doing *X*." For example, starving one's child by simply omitting to feed it is killing it. The deeper issue is what counts as a normal state of the world such that an omission leading to a death is a killing. Special duties and related expectations are not sufficient. The lifeguard who fails an obligation to make an easy rescue has not killed the swimmer. The omission must count for us as the initiation of the threat of death.

[48] This point was suggested by an editor of this JOURNAL.

If I am right about this "privilege of innocence," i.e., the special protection due those who are unthreatened and unthreatening, then double effect's distinction between intention and mere foresight lacks ultimate significance. This conclusion is unsettling since it is in fact double effect (or some closely related distinction) that appears to account for the quite powerful thought that what is particularly wrong is to harm some *in order* to aid others, as opposed to harming some as a collateral but unintended effect of aiding others. Discounting the distinction made by double effect is to discount what is intuitively and historically significant in deontology. One might attempt to incorporate double effect in a patient-based deontology by contending that the privilege of innocence is embodied in a right not to be intentionally harmed (or in the Kantian variant of double effect, a right not to be used merely as a means). There would be no comparable right not to be foreseeably harmed (or knowingly harmed in a way where one was not used). My belief now is that such a view would be patient-based only in a secondary sense. The primary wrong would be "aiming" in a certain way at the harm one caused. The right of someone not to be harmed in that way would be derivative.[49]

SOME CONCLUDING REMARKS

My aim was to sketch a patient-based defense of deontology to answer Scheffler's challenge that any defense will violate the apparently plausible goal that, other things being equal, one should act so as to minimize what (in an impersonal sense) is morally objectionable. The defense is based on giving moral weight to someone's status at a moment of not being significantly threatened with harm. We embody this recognition by ascribing rights, for example, the right not to be killed. What is salient about rights is the "non-trade-off principle," i.e., minimizing right's violations is not a sufficient reason for violating a right. The principle follows directly from recognizing the special protection due the person not threatened. That protection entails that it would be wrong for me to kill *X* to nullify my prior attempt to kill *Y*. The extent of this protection is admittedly unsettled, but at least we can argue that there is no paradox of the sort Scheffler raises about a restriction against killing one person though that would save even two or more others.

---

[49] Nagel, defending a strong prohibition against intentional harm, mentions the "special outrage" the victim feels in being "deliberately harmed" for the ("even greater") good of others. This suggests he thinks the right involved is the right not to be harmed as an instrument to achieve another's ends. There would be no comparable right not to be unintentionally (though foreseeably) harmed. See *The View From Nowhere*, p. 184. Yet one can develop examples, consistent with the strictures of double effect, where concern for the victim vanishes entirely as a consideration. Imagine a case where, given the good to be achieved, I believe (as a subscriber to double effect) that it is permitted to foreseeably, painfully, but unintentionally kill Smith. Smith requests that, instead, I intentionally and painlessly kill him (a possibility he discovered) in order to achieve the same good. As a subscriber to double effect, I refuse, and kill him painfully. The thought experiment, though admittedly strange, suggests that double effect's conception of morality as the way one's will relates to what one does is perhaps not at bottom primarily concerned with either welfare or rights. It may be the only truly agent-centered morality.

And, in addition, I noted that this defense of deontology is in accord with the "principle of maximizing rationality," MR, given Scheffler's quite general formulation of that principle; for an agent can have the goal at each moment of honoring the special protection embodied in the right of a person not to be harmed. And she best achieves the goal by not harming the person. Reason is satisfied.

Of course, no defense of rights, whether Kantian or contractarian or whatever, may prove successful. In particular, the notion I have sketched of "inertia" in the moral domain may not pan out. But the strength of the belief that it is wrong to cause harm even to minimize one's own harming suggests that moral concern is with the person unthreatened at that moment and not with whether we, rather than others (or natural forces), cause the harm. It is that concern which motivates the ascription of rights. I would note, as well, and somewhat in passing, that nothing in Scheffler's challenge appears to disallow the ascription of rights and therefore of right's violations. His point, even here, is that, if the violation of a right is objectionable regardless of who causes it, it seems paradoxical to forbid a violation whose sole effect is to minimize violations. But this raises the specter of paradox at a much too abstract and therefore misleading level. If a conception of rights is defensible, there is nothing unreasonable about a patient-based moral point of view which denies that, implicit in thinking right's violations to be objectionable, is a demand to minimize violations overall.

RICHARD BROOK

Bloomsburg University

BOOK REVIEWS

Studies in the Way of Words. PAUL GRICE. Cambridge: Harvard University Press, 1989. 394 p. Cloth $32.50.

This collection of essays brings together Paul Grice's writings on the problems of meaning broadly conceived. Ten of them have previously appeared in print; ten have not. His William James Lectures, presented in 1967 and revised in 1987, are here published in their entirety for the first time.

In various places, Grice indicates how his thoughts about problems concerning meaning developed. He describes with warmth and nostalgia his participation in the movement known as ordinary-language philosophy (OLP) which arose at Oxford after World War II. He further tells us that he never gave up the idea that ordinary language is both a suitable topic for philosophical investigation and a genuine source of philosophical insight. But even if Grice was one of the chief and most gifted practitioners of OLP, he was also, perhaps, its most penetrating critic. Following the maxim that the meaning of an expression is its use in the language, it was common among OL philosophers to point out (almost always correctly) that some philosopher had employed an expression in a way in which it is not ordinarily employed. For example, J. L. Austin would note such things as that in a dense fog it might be natural and correct to say that it looks like there is a pig over there or, perhaps, there appears to be a pig over there, but on a perfectly clear day, with the pig plainly in view, it would surely be inappropriate, odd, a misuse of language to speak in either of these ways. This and similar remarks about how we actually employ certain expressions were supposed to reveal fundamental flaws in various traditional theories of perception, in particular, those which attempted to ground empirical judgments in judgments about appearances.

Grice smelled a rat. In a presentation before the Aristotelian Society (1961), he undertook a defense of a causal theory of perception, at least against criticisms of this kind. He maintained that conversational oddness *by itself* did not show that an utterance is either meaningless or false. He pointed out that conversational oddness can arise for various reasons, one being that the speaker's use of an expression may *suggest* or *imply* something false or misleading. This, Grice noted, is compatible with the truth and, of course, the meaningfulness of what the speaker has actual said. He then went on to show in close detail that a causal theory of perception could be defended against OLP criticisms along just these lines.

0022-362X/91/8804/213–9

This contrast between what a speaker *actually says* and what the saying of it *conversationally implicates* (a label Grice introduced later), once pointed out, seems immediately plausible. In fact, Grice could have formulated his defense of the causal theory of perception quite effectively without going beyond this intuitive level. It was, however, completely uncharacteristic of Grice to let matters remain at an intuitive level. Since his argument depended upon a contrast between what the speaker actually says (the literal meaning) and what the saying of it implies (its conversational implicatures), he felt obliged to provide a detailed theoretical account for each of them. Both of these tasks were pursued in Grice's William James Lectures.

"Logic and Conversation" (Essay 2) was an underground classic before it first appeared in print in 1975. In this essay, Grice developed his account of conversational implicature, basing it on the idea that linguistic exchanges are instances of shared rational purposive activity governed (for the most part, at least) by what he called the cooperative principle and a set of more specific maxims that fall under it. Although there is room for disagreement concerning details, Grice's general approach to these pragmatic features of language is almost certainly correct, since it is able to give plausible and detailed accounts of indirect uses of language over a very wide range of cases. This part of his program is money in the philosophical bank.

As noted, Grice's attention was first drawn to the phenomenon of conversational implicature in his efforts to defend a causal theory of perception against a style of criticism common among OL philosophers. In fact, Grice constantly set himself apart from other OL philosophers by defending at least the intelligibility of many traditional philosophical positions that OL philosophers were particularly interested in rejecting. In various places in his works, he invokes the notion of conversational implicature to defend a causal theory of knowing (Essay 3), a sophisticated version of a causal theory of meaning (throughout), something like a correspondence theory of truth (Essay 3), and Bertrand Russell's theory of definite descriptions (Essay 17). It seemed to be Grice's particular calling to resuscitate philosophical positions that OLP had claimed to have buried.

In the third of his William James lectures, "Indicative Conditionals," Grice employs this method of philosophical resuscitation in an effort to show, among other things, that the logician's notion of material implication specifies the correct semantic content of ordinary 'If *p*, *q*' constructions. An important limitation on this discussion is that Grice does not take this approach to 'If *p*, *then q*', which he takes to mean 'Supposing *p*, (then) *q*' (63). Excluding such indicative conditionals that, as he thinks, express reasoning under a supposition, Grice invokes the notion of conversational implicature to defend the material implication account of indicative conditionals.

Grice's defense is most effective in dealing with problems that turn upon the so-called paradoxes of implication. "Surely it is an abuse of language," the OL philosopher will say, "to employ an indicative conditional just because its consequent is true or its antecedent false." Grice acknowledges that it would indeed be odd to use an indicative conditional when one knows either of these things, but this oddness, he argues, arises from using the weak construction 'If *p*, *q*' when we are in a position to use the stronger constructions 'not *p*' or '*q*'. Grice calls this a violation of the rule of quantity. It is this violation of the rule of quantity which accounts for the conversational oddity of employing an indicative conditional in such a context, and the oddity, thus explained, is compatible with the truth and meaningfulness of what has been said. This seems right.

This, however, is not the end of the story, for Grice encountered deeper difficulties in dealing with the denials of indicative conditionals. As Charles L. Stevenson noted some time ago, if we treat indicative conditionals as material conditionals, we can pass from the (apparent) truth 'It is not the case that if God exists, there is evil in the world' to the conclusion that God exists and there is no evil in the world. A successful theology and a theodicy should be harder work. Grice attempts to deal with the problems that arise from the denials of conditionals by introducing a bracketing device intended to control the scope of the negation sign. Sometimes the force of denying a conditional is given by the denial of the material conditional and here no component is bracketed. Grice gives clear and unforced examples of conditionals of this kind. Sometimes the antecedent is protected from negation by brackets—perhaps because it is common ground between the parties in the conversational exchange—and then the denial attaches only to the consequent. At other times, the conditional is not denied at all, and the denial is aimed at some conversational implicature of the use of the conditional. In effect, the entire conditional is bracketed, though Grice did not exactly put it this way (see 81–2).

Grice's strategy, then, for defending the thesis that the semantic content of 'If *p*, *q*' constructions is given in the counterpart material conditional was to use, in tandem, the notions of conversational implicature and bracketing. Later (in Essay 17) Grice uses the same double strategy in defending Russell's theory of definite descriptions against criticisms leveled against it by Peter Strawson. It must be said that Grice does not work out his theory of bracketing in full detail, and it remains to be seen whether it can be made to work.

The first four William James Lectures were primarily dedicated to developing the theory of conversational implicature and then applying it to certain philosophical issues. The second main task of his William James Lectures was to develop a theory that would explain speaker's or utterer's meaning using "the notion of intention in a nested kind of way" (283). Grice's interest in this program in fact antedated his concern with conversational implicature. "Meaning," where this theory was first sketched, was published in 1957, and apparently given as a lecture almost ten years earlier in 1948. About this aspect of his philosophizing, Grice remarks that "it has given me more trouble, and it has also engendered more heat" (349) than any other.

In the early stages of developing his position, Grice drew a distinction between natural meaning and non-natural meaning—the difference, for example, between smoke meaning fire and John meaning his uncle when he uses the expression 'the old codger'. (Here other philosophers have commonly drawn a distinction between natural signs and conventional signs, but, for reasons I shall examine in a moment, Grice did not want to draw the distinction in this way.) He presented what we might call his core account of non-natural meaning (meaning$_{NN}$) in these words:

> "A meant$_{NN}$ something by x" is roughly equivalent to "A uttered x with the intention of inducing a belief by means of the recognition of this intention" (219).

Starting with this relatively simple specification of meaning$_{NN}$, Grice's position passed through an extraordinary series of refinements and elaborations. Some of these elaborations were the result of technical demands, for example, to show that this simple formula could be extended to deal with the full variety of speech acts as they arise in various linguistic settings. I shall not examine such details here. The second pressure for elaboration came from the need to respond to counterexamples. Here there is only space to say a few programmatic things about these counterexamples.

Schematically, the original counterexamples to Grice's core account of meaning$_{NN}$ seemed to work in the following way: if the Gricean definition of meaning$_{NN}$ involves the mutual recognition of intentions up to some level L, then it always seems possible to produce a counterexample to the definition by introducing a bad or sneaky intention at level $L + 1$. In Grice's words, "the hearer is represented, at some level or other of embedding, as having . . . a misapprehension with regard to what is expected of him" (302–3). A quick way of dealing with counterexamples of this kind would be to add a negative clause to the definition proscribing the use of such sneaky intentions. In a sense, Grice does take this way out of his difficulty, but only after addressing a much deeper difficulty.

Setting aside problems with particular counterexamples, in "Meaning Revisited" (Essay 18), Grice argues that his account of meaning$_{\text{NN}}$ does, in fact, commit him to saying that, in fully or optimally meaning something, one is committed to the unlimited nesting of intentions. Then, instead of attempting to revise his position to avoid this consequence, he actually embraces it, thus giving his position a remarkable new turn. Grice arrived at this position from the following line of reflection:

> One might start with the idea that when some speaker S utters some sentence to a hearer A, meaning by it that p, he does this wanting A to think 'p'. That is, at stage one we have "S wants A to think 'p,' " where p represents the content of A's thought or intended thought. However, for reasons that came up long ago, having to do with the distinction between natural and nonnatural meaning, we cannot stop at stage one. We have to proceed to stage two, at which we get "S wants A to think 'p, on the strength of the idea that S wants A to think "p" ' "; and so on (300).

This is obscure, but, very roughly, *S*'s intention that his intention be recognized cannot itself be conveyed merely as a natural sign. *S*'s intention that his intention be recognized is also something that *S* intends to be recognized. It seems, then, that speakers are always in need of one more intention in order to mean fully what they intend to mean. As Grice puts it:

> what we specify A as intended to think of as the reason for thinking "p" is always one stage behind what the speaker envisages as the reason why he wants A to think "p" (300).

It follows from this that the conditions necessary for someone to mean$_{\text{NN}}$ something are logically impossible to fulfill (see 301). It seems, then, that, on Grice's account of meaning, meaning something is something a speaker is never able to pull off.

Grice's radical proposal for dealing with this difficulty is to suggest that on certain occasions "it is legitimate, or perhaps even mandatory, for us to *deem* [the speaker] to satisfy this unfulfillable condition [of endlessly nested intentions]" (emphasis added, 302). But what is deeming? And why countenance it? Grice does not answer these questions in detail. He suggests an interesting parallel with a problem in the theory of knowledge. If to know something, you cannot be wrong about it, then, according to Grice, it will be correct to say that "*strictly speaking* the only things that can be known are necessary truths" (301). He then adds:

> However, that does not restrict us to supposing that people who talk about knowing other things are using the word "know" improperly: all it requires is that there should be some license to apply the word non-strictly to things which in some way approach or approximate to the ideal cases (301–2).

With this suggestion, the burden of the argument shifts to the evaluative notion of an *ideal* case. In fact, in this volume, Grice says little about his views concerning values except for indicating, in various places, his commitment to the existence of objective values and stating, very darkly, that an appeal to objective values will be necessary "to exhibit the anatomy of speaker meaning itself" (299). This presumably is related to specifying the conditions under which it is legitimate or even mandatory to deem a condition fulfilled when, in fact, we know it to be unfulfillable. These reflections also provide a rational ground for ruling out counterexamples that involve sneaky intentions. Such sneakiness in intentions provides us with a positive reason for not deeming them ideal.

Grice's reflections on these matters are presented in a section of "Meaning Revisited" candidly entitled "The Mystery Package." The package will remain a mystery until Grice's unpublished writings on values and teleology find their way into print. Members of the Berkelischer Kreis indicate that progress is being made on this front.

I have said nothing about the final step in Grice's account of meaning: the attempt to explain word meaning and sentence meaning in terms of utterer's meaning. Roughly: (1) utterer's meaning is to be explained in psychological terms, primarily, with reference to nested intentions, and (2) sentence meaning and word meaning are to be explained in terms of utterer's meaning. Thus, in the end, complex word and sentence meanings are grounded in complex psychological states. One difficulty with this view is that certain complex intentions, as Wittgenstein has claimed, seem available only to creatures who *already* possess a language. A cat may stalk a bird with the intention of catching it, but it seems that a cat could not intend to stalk a bird a week from Tuesday. The later intention, it seems, could only be formed by a creature already in possession of a language. Similarly, the complex nested intentions that Grice invokes to explain utterer's meaning would seem to be unavailable to a creature who did not already possess a language.

I do not think that these considerations provide a knock-down argument against Grice's account of utterer's meaning and its relationship to word and sentence meaning. It might be possible to tell a reasonable story showing how prelinguistic intentions (of the kind that animals seem to have) can somehow undergird a system of primitive conventions that, through bootstrapping, can develop into the rich system of linguistic conventions that we actually use in saying things meaningful. Grice actually tells a story something like this in "Meaning Revisited," but unfortunately, right from the start, the intentions that Grice invokes seem to be those which only someone who already commands a conventional language could form. Thus, since he has not shown how semantic bootstrapping can actually take place, Grice's plausibility argument for his general program seems not to have been made out.

I have concentrated on two aspects in Grice's philosophical career: (i) his development of a theory of conversational implicature grounded in a rich conception of language as a rational, cooperative, goal-directed activity, and (ii) his attempt to develop a general theory of meaning based on the fundamental notion of intentions taken in a "nested kind of way." It seems to me that the first program has been largely successful and has had a fundamental impact upon the way philosophy has been practiced in the second half of this century. The second program has been less successful and may, in the end, seem wrongly conceived. But both programs reveal a mind constantly trying to get to the bottom of things. Paul Grice was a philosopher incapable of superficiality.

ROBERT J. FOGELIN

Dartmouth College

Against Method (revised edition). PAUL FEYERABEND. New York: Verso, 1988. vii, 296 p. Cloth $50.00, paper $16.95.
Farewell to Reason. PAUL FEYERABEND. New York: Verso, 1987. 327 p. Cloth $49.95, paper $15.95.

Unlike most revised editions, the new *Against Method* should be noticed. Most of the original[1] is still there, although if you scanned

---

[1] *Against Method* (London: New Left Books, 1975). The true originals of the book are more like "Problems of Empiricism," in *Beyond the Edge of Certainty*, R.

0022-362X/91/8804/219–23

only the indices of the two editions you might not guess that they are versions of the same book. Ephemeral polemics with minor figures have been removed, and the showpiece confrontation with Imre Lakatos's "Methodology of Scientific Research Programmes" has been dissolved. The show-off catch phrases that everyone remembers—anarchy/dada, "anything goes"—have been demoted, and a solemn introduction written for Chinese makes plain what the casual reader of the first edition often missed, namely, what the book was about. "*The events, procedures and results that constitute the sciences have no common structure*"; far-reaching changes of outlook in the sciences have happened in different ways, and given different impulses to future research (1). The book has an argument by powerful examples that make its central thesis plausible.

The new last chapter is an engaging intellectual autobiography with an important reminder of some of the people who (in Feyerabend's self-conception) profoundly affected his thinking: G. E. M. Anscombe was a primary source, as were Philipp Franck, Victor Kraft, and Felix Ehrenhaft. And it was C. F. von Weizsäcker who put a stop to Feyerabend's alienated assaults on quantum physics. The autobiography ends with an honorable confession of regret. He once rather bombastically thought that his relativistic philosophy could serve the students that Berkeley brought in from the ghettos and barrios of California. Now: "it is conceited to assume that one has solutions for people whose lives one does not share and whose problems one does not know. . . . My own view . . . was still a view, an abstract fancy I had invented and now tried to sell without having shared even an ounce of the lives of the receivers. This I now regard as insufferable conceit."

The heading for the autobiography states that, above all, the motive force behind Feyerabend's work is anger: anger at intellectual conceit, at "treacly phrases" of intellectuals, at destruction of, interference in, or control of other people's lives. As for those who pro-

---

Colodny, ed. (Englewood Cliffs, NJ: Prentice-Hall, 1965), pp. 145–260; and "Problems of Empiricism Part II," in *The Nature and Function of Scientific Theories*, R. Colodny, ed. (Pittsburgh: University Press, 1971), pp. 275–354; neither of which occurs in *Problems of Empiricism: Philosophical Papers*, Vol. 2 (New York: Cambridge, 1981), but both of which are largely retained in the revised edition of *Against Method*. There is also a German version whose third edition is different from any English one, and which embodies the final, eponymous, chapter of *Farewell to Reason*; cf. *Wider den Methodenzwang* (Frankfurt: Suhrkamp, 1986).

claim or fear that there is "cultural chaos" that "pervades the essence of our culture," they are speaking about the rantings of a handful of highbrows. We should rather fear what François Jacob calls the threat of "monotony and dullness."

Feyerabend's considered philosophy is strictly critical, constructive only in urging people to learn from each other but otherwise to live and let live. It rejects without sadness all claims to permanence or foundations. His arguments are ad hominem, and his metaphilosophy is a commitment to ad hominem criticism. So he can cheerfully tell the mandarins who bewail "cultural chaos" to stuff it and watch the Superbowl and *Dynasty* to see how little chaos there is in their culture—and then on other pages regret how one technology and its values can sweep aside all others, empowering *Dallas* and (Association) football not only in Rome but also in Burkino Faso. He undoubtedly thinks that relativism is some sort of prophylactic on both fronts.

The longest essay in *Farewell to Reason,* "Notes on Relativism," does not urge relativism as a single monolithic doctrine. Instead, a series of eleven loosely relativist positions is rehearsed. All are presented as reactions against the militant claims of capital *R* Reason. He has no objection to being reasonable or objective, by whatever immediate canons we may have to hand. He is against the notion that there is some ultimate objective and rational way of finding out what is true. "The ideas of truth, reality and rationality make excellent practical sense"; Feyerabend opposes only the metaphysical constructions therefrom, the One Truth, the One Reality, the One Rational Method.

The eleven subsequent essays in the collection vary in interest. They are best when they present a simple figure as complex, as in the discussion of Ernst Mach, and worst when they present a complex figure as simple, as in the essay on Karl Popper. The deflating of Popper is irrelevant to American culture, where Popper is of no significance, but elsewhere he is the most widely read philosopher of the twentieth century. Hence the placing of Popper within the problematic of logical positivism may be of value, and there is a nice diagnosis of why so many scientists took him so seriously. There is a very tidy essay on Aristotle's theory of mathematics, and a brisk, accurate, rebuttal of Hilary Putnam's flippant contention that Feyerabend's idea of incommensurability and relativism are self-refuting. The concluding chapter, called "Farewell to Reason," is based on a reply to two German volumes of essays about Feyerabend. Like most replies of the "Schilpp" sort, that reply did not come off well in the first place; it is not satisfying on its own as the conclusion to this book.[2]

Undoubtedly, these essays will be read chiefly to find out Feyerabend's latest thoughts about relativism. I believe he says one thing that is completely wrong, one that is happily belied by his own practice. He says that the kind of relativism he presents is "about human relations. It deals with problems that arise when different cultures, or individuals with different habits and tastes, collide." But he is not talking about the collision of cultures in general. Despite the occasional third-hand references to exotic peoples, his relativism is "ours" and is our Greek heritage. Look at the proper names in the index. Bohr, Einstein, Galileo, Mach, and Popper have big entries, but after that, it is Greeks all the way. That is as it should be. Only "Westerners" have a problem about relativism, and we know where it comes from. *Farewell to Reason* may be the closest we will get to a book Feyerabend has never quite written, an assault on the entire tradition engendered by those imperial Greek rationalists who had us replace magic by mathematics.

Relativism (as opposed to what Feyerabend rightly calls opportunism—learn from the others what is handy) does not arise automatically from cultural confrontation. Peoples have been crossing paths since the beginning, with consequences that range from ruthless genocide, through trade-offs and the division of labor, to straightforward indifference. Most histories, be they of the Pacific Northwest of this continent or 3000 years of Persia, have the usual mess of slavery, pillage, marauders, and mayhem interspersed with quite a lot of more peaceful commerce. Just like "our" history. Many peoples, and certainly Westerners, have been confident of their own ways, often to the point of extinguishing others. But it is only "we" who have been relativists, because of our Greeks who articulated the abstract concepts of truth, objectivity, and reason.

Travelers have always known about and been amused or disgusted by the pollution rules of other people, but it was Herodotus who (on Feyerabend's undisputed telling) first drew the relativist conclusion. Darius set up Herodotus by introducing Greeks to people who ate their dead parents, and who thought that the Greek practice of cremation was heinous. Darius was merely (I suppose) mocking the

---

[2] H. P. Duerr, ed., *Versuchungen*, 2 vols. (Frankfurt: Suhrkamp, 1980, 1981). An English translation of the essays and reply is being prepared by G. Munévar, *Beyond Reason: Essays on the Philosophy of Paul Feyerabend* (forthcoming in the Boston Studies in the Philosophy of Science Series).

Greek objectivity mongerers, but Herodotus was fully serious in the way that only a Western rationalist can be. What if my enthusiasm for science is no less a matter of custom than my revulsion at the thought of eating my dad's corpse? This tale (which Feyerabend has used for decades) stands for the intense Greek debates about objectivity and relativism which their successors have sometimes muted but never forgotten. And it was not ethical relativism that came first, at least on the testimony of Aristotle, but qualms about this new conception, objective knowledge.

To repeat, relativism is not about "what happens when different cultures . . . collide." It is about what happens when "our" apparently unique concept of objectivity meets anyone else. Just as objectivity was located not in things but in what lay behind things, or in the reasons traversed and surveyed by the mind, so the doubts that follow the examination of other knowledges thrust us inward upon ourselves. "Anything goes," Feyerabend confessed in the preface to the new *Against Method,* was "the terrified exclamation of a rationalist who takes a closer look at history." And the eleven relativist theses of "Notes on Relativism" are the rationalist whimpers that follow this cry of terror.

IAN HACKING

University of Toronto

NEW BOOKS: ANTHOLOGIES

ADAMS, PARVEEN, and ELIZABETH COWIE, eds.: *The Woman in Question:* m/f. Cambridge: MIT, 1990. 383 p. Cloth $25.00.

BARRETT, ROBERT B., and ROGER F. GIBSON, eds.: *Perspectives on Quine.* Cambridge: Blackwell, 1990. xxiii, 358 p. Cloth $45.00.

BRICKER, PHILLIP, and R. I. G. HUGHES, eds.: *Philosophical Perspectives on Newtonian Science.* Cambridge: MIT, 1990. vii, 248 p. Cloth $34.95.

COHEN, PHILIP R., JERRY MORGAN, and MARTHA E. POLLACK, eds.: *Intentions in Communication.* Cambridge: MIT, 1990. xi, 508 p. Cloth $45.00.

EVERSON, STEPHEN, ed.: *Companions to Ancient Thought 1: Epistemology.* New York: Cambridge, 1990. viii, 288 p. Cloth $54.50, paper $16.95.

FINE, ARTHUR, MICKEY FORBES, and LINDA WESSELS, eds.: *PSA 1990, Volume One.* East Lansing, MI: Philosophy of Science Association, 1990. xxix, 599 p. Cloth $22.00, paper $20.00.

FRENCH, PETER A., THEODORE E. UEHLING, JR., and HOWARD K. WETTSTEIN, eds.: *Midwest Studies in Philosophy: Volume XV: The Philosophy of the Human Sciences.* Notre Dame: University Press, 1990. 464 p. Cloth $49.95, paper $21.95.

GEORGE, ALEXANDER, ed.: *Reflections on Chomsky*. Cambridge: Blackwell, 1990. xii, 271 p. Cloth $45.00.

GILL, CHRISTOPHER, ed.: *The Person and the Human Mind: Issues in Ancient and Modern Philosophy*. New York: Oxford, 1990. x, 284 p. Cloth $55.00.

GOLDBERG, DAVID THEO, ed.: *Anatomy of Racism*. Minneapolis: Minnesota UP, 1990. xxiii, 357 p. Cloth $39.50, paper $14.95.

GOODIN, ROBERT E., and ANDREW REEVE, eds.: *Liberal Neutrality*. New York: Routledge, 1990. iv, 219 p. Cloth $49.95.

JACOBUS, MARY, EVELYN FOX KELLER, and SALLY SHUTTLEWORTH, eds.: *Body/Politics: Women and the Discourses of Science*. New York: Routledge, 1990. vii, 107 p. Cloth $39.50, paper $13.95.

KELLY, MICHAEL, ed.: *Hermeneutics and Critical Theory in Ethics and Politics*. Cambridge: MIT, 1990. xi + 285 p. Paper $13.95.

KEMP, T. PETER, and DAVID RASMUSSEN, eds.: *The Narrative Path: The Later Works of Paul Ricoeur*. Cambridge: MIT, 1990. viii, 121 p. Paper $10.95.

KOELB, CLAYTON, ed.: *Nietzsche as Postmodernist: Essays Pro and Contra*. Albany: SUNY, 1990. vii, 350 p. Cloth $59.50, paper $19.95.

KYBURG, HENRY E., JR. et al., eds.: *Knowledge, Representation, and Defeasible Reasoning*. Boston: Kluwer, 1990. xxii, 423 p. Cloth $99.00.

LYCAN, WILLIAM G., ed.: *Mind and Cognition: A Reader*. Cambridge: Blackwell, 1990. ix, 683 p. Paper $19.95.

MACDONALD, SCOTT, ed.: *Being and Goodness: The Concept of the Good in Metaphysics and Philosophical Theology*. Ithaca, NY: Cornell, 1991. ix, 328 p. Cloth $43.95, paper $14.95.

OSHERSON, DANIEL N., et al., eds.: *An Invitation to Cognitive Science: Volumes 1–3*. Cambridge: MIT, 1990. xix, 273 p.; x, 356 p.; x, 308 p. Each cloth $37.50, paper $18.95.

PETERSON, GRETHE B., ed.: *The Tanner Lectures on Human Values, Volume XI, 1990*. Salt Lake City: Utah UP, 1990. ix, 578 p. Cloth $25.00.

SAVAGE, C. WADE, ed.: *Scientific Theories*. Minneapolis: Minnesota UP, 1990. ix, 429 p. Cloth $30.00.

SCHNEEWIND, J. B., ed.: *Moral Philosophy from Montaigne to Kant: An Anthology: Volume I and Volume II*. New York: Cambridge, 1990. xxi, 664 p. Vol. I: Cloth $49.50, paper $14.95. Vol. II: Cloth $49.50, paper $13.95.

SEREQUEBERHAN, TSENAY, ed.: *African Philosophy: The Essential Readings*. New York: Paragon, 1991. xxii, 250 p. Paper $13.95.

SOLOMON, ROBERT, and KATHLEEN HIGGINS, eds.: *Reading Nietzsche*. New York: Oxford, 1990. xiv, 251 p. Cloth $32.50, paper $12.95.

SORABJI, RICHARD, ed.: *Aristotle Transformed: The Ancient Commentators and their Influence*. Ithaca: Cornell, 1990. x, 545 p. Cloth $65.00.

WOLTER, ALLAN B.: *The Philosophical Theology of John Duns Scotus*. MARILYN MCCORD ADAMS, ed.: Ithaca: Cornell, 1990. ix, 356 p. Cloth $47.50.

WREN, THOMAS E., ed.: *The Moral Domain: Essays in the Ongoing Discussion between Philosophy and the Social Sciences*. Cambridge: MIT, 1990. xxix, 414 p. Cloth $35.00.

YOURGRAU, PALLE, ed.: *Demonstratives*. New York: Oxford, 1990. vi, 275 p. Cloth $39.95, paper $12.95.

THE JOURNAL OF PHILOSOPHY

VOLUME LXXXVIII, NO. 5, MAY 1991

THE WHAT AND THE HOW*

The truth about Nastassja is this: she is a human being; she is also an actress, herself born to a famous actor. Discovered at a tender age by Roman Polanski, she was to be embroiled in much controversy that soon turned to fame. The French film critics, otherwise notorious for their iciness, describe her, breathlessly, as the long awaited heir, in both style and substance, to the legendary Ingrid Bergman. "Divine and sublime," wrote one of them reviewing her last appearance in the Cannes festival; "Sublime and divine," echoed another. In a more philosophical mood, and lest it be Griceanly inferred that the traits following the above "also" are meant to be modally contingent, I note that Nastassja is a twentieth-century figure, is not yet thirty years old, is Caucasian, and is of German descent; all are modally essential to her, if current essentialist doctrines are to be trusted.

---

* Work on the idea of logical truth which I had undertaken in the mid-1980s was goaded by *Tractatus* 5.552: "Logic is prior to the How, not to the What." Alas, I never really got, in that work, to reflecting on the distinction alluded to by this line, let alone to why logic would be prior to the one but not the other.

I have come upon the same distinction through a different lens, work on Descartes's discerning of mind and body. While attempting to reconstruct an argument for their discernibility that is neither from how we know the interested parties nor from alleged modal possibilities open to them, the fundamentality of the purported conceptual distinction grew on me. Sometime in 1987, I decided to compose an elementary exposition of this distinction, as I saw it. There resulted the present work.

I owe many thanks to Roger Albritton, Keith Donnellan, Y. Gutgeld, W. D. Hart, A. Hsu, and P. Leonardi. Especially helpful were David Braun and David Kaplan who, through close readings of early drafts, provided a constructive proof of how the intended distinction may be mistaken for others.

I refer to Bertrand Russell's *Introduction to Mathematical Philosophy* (London: Allen Unwin, 1919) as IMP.

0022-362X/91/8805/225-44

Truths about Nastassja they all are. But there is a difference in what *makes* them, in the actual world, into such truths: that Nastassja is human is true solely in virtue of *what* she is; the rest are made true not by what but merely by *how* she is. This notion, truth in actuality solely in virtue of what the subject is, involves, at least at first blush, neither mention of how we know the truth nor reference to its stability in counterfactual worlds. Later blushes are to confirm this: we have on our hands an autonomous notion of distinguished truth. I call truths solely in virtue of what their subject is *primal truths*. The elucidation of this notion is what this paper is about.

I. THE INTUITIVE BASIS OF PRIMAL TRUTH

I.1 The coherent-subtraction test. Try to subtract from the above profile the fact that Nastassja is actually human, i.e., human in the actual world. For that matter, try to subtract from Nanga Prabat's profile its being a mountain, from Venus's its being a planet, from the Mona Lisa's its being a painting, from the Notre Dame's its being a cathedral. In the same vein, set yourself to tinker with the profiles of slightly more "abstract" subjects, e.g., the species whale and snail, the substances gold and water. Deprive the first of its being mammalian, the second of its being molluscan, the third of its being an element, and the fourth of its being a compound. All the attempted subtractions seem incoherent. Why?

Although Nastassja is my favorite subject, allow me, in view of the delicacy of this particular case, to begin with a more robust paradigm. Let it be the Nanga Prabat. In subtracting a trait from the profile of such a particular subject, e.g., 'being snow capped', we must strike a delicate balance between (i) merrily stripping away the subject of its predicates and (ii) preserving its integrity, viz., the sense that it is still this subject from whose profile we subtract. We can subtract with impunity snow-cappedness from Nanga Prabat's profile because, while we so falsify the record as to *how* this subject is, our sense of *what* is thus misdescribed remains intact. Intact it remains because our presuppositions, those which cement the very sense that it is this particular item with which we tinker, have been vigilantly kept out of the subtractive melee.

Alas, sometimes greed gets the better of us and our subtraction experiments violate the very presuppositions that make them possible. Attend again to the Nanga Prabat. Whatever liberties we are about to allow ourselves, if that worldly peak itself is to be the subject of our subtraction experiment, we must not tinker with its being a mountain. But then what of one who claims that 'being a mountain' is subtractible, too, offering as his "proof" the coherence of a supposition, a "short story" if you will, in which N.P. itself bears proudly the predicate 'not a mountain'?

The predicating hand destroys what the one handling the subject took so much care to protect. It is shamelessly supposed of our subject, a mountain if nothing else, that it is not one. Inevitably, the suspicion grows that our subject is not Nanga Prabat but merely an "arbitrary" entity, with no identity, a close kin to the philosopher's mysterious "bare particular." And so, with our most fundamental presupposition violated, the aboutness of the subtraction experiment slips away. Gone is the sense that it is still about the real Nanga Prabat, not an invented philosophical fiction, that we suppose.

Call the question—Is it coherent to subtract from *x*'s actual profile the trait *F*?—*the coherent-subtraction test.* Four basic guidelines about how it should be read are in order. First, I should like it noted that it is only a test, viz., a method for identifying primal truths, rather than an analysis of the notion or an explanation of what *makes* a truth primal. In particular, the dependence on the act of subtraction and on what the subtractor(s) can or cannot mentally subtract should not be taken as constitutive of *x*'s whatness. In agentless primordial times, 'mountain' already did and 'snow capped' already did not pertain to what the Nanga Prabat was (and is).[1]

Coherent subtraction is only a test. What is it a test *for*? The test seeks to discover whether a given trait *F* in *x*'s profile pertains to its very whatness. Equally important, however, is to note what it is *not* a test for; the test leaves us free of a variety of (i) semantical, (ii) epistemic, and (iii) modal involvements. Thus, the second guideline is that the present metaphysical speculations about the nature of things, e.g., Nastassja, should not be confused with semantical hypotheses about their names, e.g., 'Nastassja'. In claiming that 'human' is a primal trait of *N*, and that the singular truth, 'that she is human', is primal, I have not claimed, e.g., like Peter Geach, that the English sentence '*N.* is human' is true in virtue of meaning alone ("analytic"), true in virtue of the supposed "connotation" of the name. Nor did I claim that, if a user of the name 'Nastassja' failed to believe its referent was human or believed it was not human, he would fail to refer to Nastassja. I am here concerned with the nature of the object itself, the character of the bond between word and object notwithstanding.

The third guideline is that the test does not involve us in epistemological conjectures regarding what we may or may not know about the subject. I take it as given that Venus is in actuality both a planet and of interest to the Russians; also, that whales are both mammalian and of interest to marine biologists. I claim that in both cases we can coherently subtract the second trait but not the first. To stress the epistemic-free character of this claim, let it be granted (though, in the mantle of an epistemologist, I would not so grant) that (a) it is consistent with what we know that Venus is not a planet and whales are not mammalian; also, the even stronger (b) that it could turn out that Venus is not a planet and whales are not mammalian. To put the matter in even less hypothetical a form, consider a sufficiently ancient counterpart of the present author, also bent on discovering the what and the how of Venus and the whales. He is on record for suggesting that 'star' and 'fish' cannot be subtracted from their respective profiles. Led astray by his fellow astronomers and biologists, our ancestor metaphysician did turn out to be wrong (only his *meta*physics was, as it were, right). This epistemic fragility of his and my pronouncements, however, is quite irrelevant to the application of the subtraction test.

---

[1] The ramifications of the idea that the subtraction test is only an identification procedure for primal truth are discussed in detail in part IV.

The epistemologist's most prized target is the ultimately true profile of, e.g., Venus. He constantly operates with a merely temporary profile, worrying whether everything with it is quite what it seems. In contrast, the ancient metaphysician and I get hold of the epistemologist's target by stipulation, i.e., we begin with Venus's actually true profile as given. By successive subtractions of its merely descriptive traits, the how it is, we hope to unravel its predescriptive traits, those which guarantee the availability of a subject, the "what," for the eventual description. As such, the subtraction-receptive traits might not be epistemically inferior. They might, if, e.g., we focus on phenomenal traits of the right kind, be known to us with perfect certitude. Conversely, the subtraction-resistant traits are not, as already stressed in the case of Venus and the whales, epistemically superior. Subtraction resistance and receptiveness are determined by the nature of the subject investigated, not by the extent and hardness of the information available to the subtractor. We discover what Venus is by subtracting; what it is determines what we can subtract.

Finally, the fourth and last guideline: the test is meant to be taken in a modal-free way. I have just emphasized that we take the way Venus actually is as given and not worry about how it might turn out. This may encourage some readers to interpret the test as trying, given the way Venus is, to "lasso" the ways it might have been or, worse, to separate, among its necessary traits, the fundamental and the derivative. This is not the way in which the test is to be taken. The modal misinterpretation being the most prenicious, a few conceptual distinctions must be introduced at this stage.

1.2 Counteractual subtraction. Our test involves what I will call *counteractual subtractions,* not modal assertions of possibility. Counteractual subtractions and modal assertions differ, as we shall discover, in many ways. At this early stage, I merely observe that the two notions differ already at the level of extension. First, modally contingent propositions might be resistant to counteractual subtractions. This much was clear to Descartes (to whose insightful employment of the apparatus of subtraction experiments I shall return). He knew all too well he was not a perfect being who existed necessarily. And yet, he insisted, in the first person, that one cannot "subtract away" one's actual existence. I am not interested here in the epistemic dividends afforded by the first-person exercise; I am rather focused on the logical point implicit in Descartes's discovery: to subtract from the profile of any actual thing, other minds or other bodies, Nastassja or the Nanga Prabat, the trait of existence is to let the lively presence of the subject flatly contradict the supposed success of the subtraction. Some readers will nonetheless protest that we say nonchalantly, "Suppose that Nastassja didn't exist. . . ." Of course we do; however, what we so do is consider a merely possible history of the world, doing so all the while with Nastassja actually on hand. We consider the world as it was twenty-seven years ago and note that, for all but the grace of God, Klaus, that most notorious of womanizers, might have left off a month earlier with his next conquest, depriving us forever of the Nastassja we did come to enjoy. We can generate this much counterfactual history *of hers* only because *she* is actually on the platform.

The converse combination of coherent, counteractual subtractions that are modally necessary is equally available. This is so if only because for each trait *F* Nastassja actually has, N. is actually *F* is modally necessary. But for many such *F*s, e.g., being an actress, we can suppose Nastassja actually *F*-less. "Actually" free cases make the same point. As modal assertions, the propositions that N. is a twentieth-century figure and that N. is human are equally necessary, if current modal metaphysics is to be our guide. In contrast, whereas counteractually subtracting the former is perfectly coherent, if not downright intriguing, dispensing with the latter leads to the irreversible loss of the subject of the whole exercise.

1.3 Existence conditions. The coherent-subtraction test is only a test. It provides a way of "lassoing" the extension of the notion of primal truth. The notion itself is not thereby analyzed. For myself, I harbor much skepticism when it comes to definitional analyses of fundamental (philosophical) notions. The aim of the remarks that follow is to supply an intuitive elucidation of the notion of primal truth. My basic idea is that, by reflecting on the coherent-subtraction test, we can distill such an elucidation.

To begin with, we need to distinguish, for any given subject, *conditions of existence* and *modes of existence*. The distinction is between traits the possession of which the subject's actual existence is conditional on, and traits the possession of which is conditional on that subject's actual existence.

More austerely, call a trait P a condition of existence for x iff it has to be actually true that, if x exists, x is P. P is a mode of existence of x iff P is a trait of x and is no condition of existence. A trait is a primal trait of x iff it is a condition of existence. A truth of the form x is P is primal iff P is a primal trait of x.

In line with the separation of counteractual subtractions and modal assertions of possibility, I note that the notion of a condition of existence is not to be confused with the modal notion of essence: conditions of existence are meant to resist counteractual-subtraction experiments, not refutations in arbitrary counterfactual worlds; they state, not what it takes for x to exist in an arbitrary possible world, but what is a *sine qua non* for x's existence in the actual world.[2]

My formulation relies on a seemingly mysterious 'has to be actually true' (though, again, arguments "by ear" do not convince me that in English it obviously means what we intend by '□'). Simple applications of the test of refutability by subtraction experiments versus refutability in an arbitrary counterfactual world should convince us (prior to much theoretical footwork that is to follow) that 'has to be actually true' differs already in extension from the 'has to' of modal necessity. As mentioned, 'N. is human', 'N. is a twentieth-century figure', and 'N. is actually an actress' are all modally necessary but only the first has to be actually true. On the other hand, there are propositions that are modally contingent but have actually to hold, if she is actually to be. 'N. exists' was already mentioned and 'If N. exists then' ('if N. is actually an actress then N. is an actress') is another such case: the proposition has to be actually true, as any other *F* we care to substitute for the trait of being an actress will verify. Nonetheless, the proposition is contingent as attested by a counterfactual world where Nastassja does not pursue a career in the movies.[3]

---

[2] There is, I believe, a traditional use in philosophy of the term 'essence' that is modal-free and is very much in the spirit of what I intend by 'condition of existence'. Perhaps even the ordinary language use of 'essential' can be so heard. In view of the dominance of the modal reading of 'essence' in current discussions, however, I shall so use it. Like remarks apply to the classical logical idea of 'necessary condition for . . .'. I hear it as a modal-free notion. But to avoid any strain *vis-à-vis* current use, I have adopted '*sine qua non* for . . .' where no alethic modality shows up.

I.4 Existence conditions and whatness. A proposition of the form '*x* is *P*' is a primal truth iff *P* articulates a condition of existence of *x*. But why, it may be asked, does the idea of condition of existence capture the primordial idea of what *x* is? In focusing on what, as opposed to how, Nastassja is, we proceed, as already suggested, by "analysis": beginning with the full profile of her traits, we go on to subtract the traits that describe her till we reach her core traits, those which guarantee the provision of a subject, the "what," for the eventual description. In contrast, in articulating existence conditions for N., we proceed by "synthesis": we pretend that N. is not around and wonder, now that we play the role of "the maker," what is absolutely a *sine qua non* for us to do if we are to synthesize her.

Of the two, the "synthetic" idea of existence condition runs deeper. It is because a trait is a condition of existence that it is predescriptive, not the other way round. It is written into the very idea of "description of a subject" that there be a subject to be described. Thus, if a trait is an existence condition, it does not belong with a description of how the given subject is but rather is part of the predescriptive stage setting: the provision of a subject, the what, whose acts and deeds are to be described. What is more, even when we engage in a deliberate attempt to specify N.'s "what" in existence-free terms (as in the above "analytic" procedure), allusion to her existence conditions is inevitable. We are to "peel off" her traits till we reach what may be termed "the critical point," the point where we do not dispense anymore with her being thus-and-such but her being at all. Her "what" is thus inextricably linked to what it takes for her to be.

---

[3] Another common misconstrual of 'has to be actually true': in oral presentations of the present material, 'has to be actually true' has, at times, been read as 'is eternally true'; this was "backed up" by some with the, allegedly "Aristotelian," idea that what I am after, viz., what Nastassja is, consists first and foremost of those features which persist with her through changes. Let me then be crystal clear about this "temporal interpretation." First at the level of the notions involved, read 'what N. is' not as 'what N. was, is, and will be'; take 'what N. is' for what it is, viz., with a tensed 'is'. Also, nothing in my corresponding, logical 'has to' alludes to truth at (all) times. It is consistent with everything I have to say that one contests the very idea of having a property at a time (and not another) and of a proposition true at one time and not another; even if this much is not contested, it is still consistent, in a "Quinean" mold, to contest the very coherence of predicating individuals, in and of themselves, regardless of how they are specified, with tensed properties; finally, even if the intelligibility of tensed predication is accepted, it is still consistent to hold the view that Nastassja never changes, that all her features are temporally eternal. This may be because she is taken to exist for, say, 70 years, but without any changes or because her life is reduced to a very short period, in which no changes occur, the subsequent acts and deeds being really of a series of connected, but distinct, individuals. So much then for the independence of the notions. In addition, I should like it noted that, even if one were to accept a "standard" view of temporal notions (I do), there would still be marked differences at the level of the extensions of the two notions, viz., eternal truth is neither sufficient nor necessary for whatness. 'Was conceived on a rainy day', 'was not really planned', and 'was born in Germany' are all eternally true of N.; if nontensed predicates are wanted, 'has a birth mark on her back' and 'has a liver' will (probably) do; none of those pertains to what she is. Also, it is not inconceivable for me that, given radical changes in the environment, taking place in a gradual way, N. could turn from being human into being some other kind of higher sentient mammal. To rule out this possibility would take some argumentation going much beyond the question of what she is (now).

II. THE KIND OF QUESTIONS "WHAT QUESTIONS" ARE: ONTOLOGY

II.1 Classification versus individuation. The history of attempts to address "what questions" is long and distinguished; it may be part of the very *image de marque* of philosophy itself (is not many an immaculate student initiated to a vivacious Socrates hobnobbing with marketplace cobblers and pondering with them the question, 'what is a shoe?'?). Reflection on the labors of such distinguished what-seekers reveals that many of them heard in the question, 'what is x?' a call not merely for existence conditions but for full identity conditions, viz., answers of the form, 'why, but it is the unique *F*', where '*F*' specifies the "real definition" of *x*.

By this, philosophers have not merely aimed at (modally) necessary and sufficient conditions for being Nastassja: witness the summary dismissal of 'the actually favorite actress of Roman Polanski' or 'the sole member of {Nastassja}'. Similarly, Richard Dedekind, G. Peano, Gottlob Frege, and Bertrand Russell, all intent on articulating what, e.g., the number 4 is, would not allow, even if they were to swallow the modern infatuation with modal essentialism, the answer, 'number that solves the map-coloring hypothesis', or even, 'number of dimensions of a Minkowski space', modally necessary and sufficient as these traits might be. What such what-seekers look for is a specification that articulates what *x*'s identity "consists in." For example, if some philosophers got their way, 'such and such continuous succession of human stages' would do in N.'s case; 'concept any first level concept falls under when four members fall under the latter', 'set of all four-membered sets', and 'successor of 3' did, *grosso modo*, the trick for the aforementioned number analysts. Responding, as they do, to 'what is *x*?' queries with individuating specifications, I shall call such what-seekers *individuation theorists.*[4]

---

[4] I have in mind a whole variety of individuation theorists. All seem to me moved, *au fond*, by "reductionist" tendencies: truths about the original subject

From the beginning of this paper, I have adopted the mantle of a classification theorist. I have presupposed, rather than argued for, my view that in articulating the whatness of many worldly subjects, as natural as Nastassja and the Nanga Prabat or as artifactual as the Mona Lisa and the Notre Dame cathedral, "defining," better called "individuating," features need not be included. This much does not deny that some trait '*F*' may uniquely apply to them, perhaps even in all possible worlds. I rather predict that such a uniquely specifying trait will merely describe how they are, perhaps how they are in all worlds, but will not divulge *what* they are. In a less presuppositional mode, I note that in response to 'what is Nastassja (Nanga Prabat, Mona Lisa)?' we do not despair and even appreciate 'a human being', 'a mountain', and 'a painting'.

In previous work, I have been inclined to challenge the individuation-oriented approach. Not here. Doubtless, the individuation theorist and I supply different answers to 'what is x?' queries; we agree, however, on the kind of question it is and, equally significant, the kind of question it is not. I have come to believe that the prize for abstracting from such "local disagreements" is double: first, a sharper, dare I say clear and distinct, perception of the kind of question 'what is *x*?' is; secondly, a better understanding of where the two pictures agree, where not, and, most important, which of these matters more.

---

can (and should) be "reduced" to truths about more "basic" items, viz., those in terms of which the individuation criteria are stated. Talk of "reduction" comes in various shades and half tones. Some will want the original subject, e.g., N., the number 4, the species whale, literally identical to certain equivalence classes (of her stages, of four-membered sets, and of animals descending from a common origin); others will avoid identity talk and seek "contextual definitions" that eliminate the merely apparent subject, in reality a "logical fiction," in favor of purely quantificational claims (Russell, at one point, thought almost everything under the sun, even the sun itself, was such a logical fiction); yet other theorists would favor the "disappearance theory" of Nastassja et alia: there is no such thing as N., we are in the grip of a "perceptual error" (perhaps Hume's theory of the self can be so read). Milder positions would preserve Nastassja's integrity. It is simply that truths about her may be taken to "supervene" on truths involving her stages (the terminology of "supervene" allows itself many different readings); the equivalence class may be taken to "represent" Nastassja; and so on.

Yet another dimension of variation among individuation theorists is the one involving extensional and intensional individuation theorists. The extensionalist correlates one subject (e.g., Nastassja) with another, extensionalized subject (e.g., the equivalence class). The intensionalist correlates a subject (e.g., Nastassja) with a predicate, viz., a condition specifying uniquely what N. is (e.g., 'continuous succession *S* of human stages'). In the text, I shall not engage in subtleties arising from the existence of various clans of individuation theorists. My remarks are framed to cover them all.

II.2 Ontological questions: the autonomy of primal truth. What kind of trait do we ascribe to a truth when we pronounce it "true solely in virtue of what its subject is"? The trait so ascribed and the question thus addressed, whether in the manner of the individuation or the classification theorist, are, I believe, rather distinct. I shall call such questions *ontological*. In what follows, their autonomy and, in particular, independence from the more familiar epistemic and modal queries, will be set forth.

Paradigmatic to epistemic questions about Nastassja is the search for propositions that have to be true because of the *way* we know that they *are* true. So read, 'it has to be true that N. is self-identical' (compare: 'that I think'), is true because of the way in which we know that the embedded claim is true. Paradigmatic to modal questions about Nastassja is the search for propositions that are true no matter which way the world might have been. So read, 'it has to be true that N. is self-identical' (compare: 'that I am the offspring of gametes *J*') is true because the embedded claim could not have been false.

Ontological investigations concern neither our knowledge of the actual world nor the truths of the nonactual ones. They concern the actual *world*, the things themselves, and they concern the *actual* world, not the merely possible others. Of course, the ontologist does not care to investigate whether N. brushes her teeth or whether the Nanga Prabat has a helicopter at its very top. The *state* of things is not the ontologist's business; the things are. He strives to articulate the make up of this world, what there is and what it is that is.

So understood, the what-seeker, individuation or classification oriented, is in search of propositions that have to be actually true, if their subject is actually to exist, if there is to be this "what." Read this way, 'it has to be true that Nastassja is self-identical' (compare: 'that, if N. actually brushes her teeth, N. brushes her teeth') is true because, if she is to come into actual being, her being self-identical is a *sine qua non*.

The autonomy of primal truth and its irreducibility to notions of distinguished truth of the classical variety calls for some additional observations, equally precious to the classification and individuation theorists. Can truth solely in virtue of what N. is be a disguised form of truth solely in virtue of our internal "concept" of N.? It seems not. Apriority, however we understand this intractable notion, is not coextensive with primal truth. We have only to recall the primality, agreed upon by the classification and individuation theorists, of, e.g., the a posteriori truths that Nastassja is human, Venus is a planet, whales are mammals, and gold is an element. Follow an individuation theorist like W. O. Quine, and truths like 'Elmer is such and such a continuous succession of tree stages' are classed definitive, true solely in virtue of what Elmer is. But as Quine himself is the first to point out, the establishment of such truths involves investigations of processes of "chemical deformation," of what is involved in being a tree, etc., all empirical matters par excellence.

A more prevalent suspicion is that, in classifying a truth as primal, we ascribe to it a modal distinction of being necessary and of so being because it specifies the subject's modal essence. The suspicion is unjustified. I have already mentioned propositions like 'N. is a twentieth-century figure', 'N. had egg E at her origin', 'N. is actually an actress'; 'The species whale originated before 1066AD', and 'gold has specific gravity larger than 1' can be added to the list. Accept current essentialism and necessary they all seem; articulators of what their subjects are they are not.[5]

Is involvement with modality at least necessary for the primalist? One has only to recall the Frege-Russell search for what the numbers are coupled with their disdain for the modal to start wondering. Let yourself be a *de re* modal skeptic: predicating of Nastassja necessary humanity, in and of herself, is meaningless. In the language of "possible worlds," you hold, whether for Quine's, Alonzo Church's, or David Lewis's reasons, that the idea of nonconventional identity across worlds is bogus. This much does not bar you from accepting that, in the actual world, 'human', and even 'continuous succession *S* of human stages' if you sympathize with the individuation theorist, pertain to what N. is, 'actress' merely to how she is. In fact, you may even radicalize your modal skepticism so much as to reject what Quine calls the (mere) "second grade involvement with modality," viz., for you necessity cannot even apply to closed sentences ("general propositions") of the object language; better confine it to the metalanguage (and read it as 'is analytic'). So be it. You can still accept that 'human (continuous succession *S* of human stages)', not 'actress', gets at what Nastassja is.[6]

---

[5] Another way of making the same point: suppose that, like some fatalists and other modal determinists, you believe that every trait N. happens to have is hers of necessity, that things could not have been otherwise. Even in this, "deterministic" setting there are necessities and there are necessities: that N. is an actress is a necessity pertaining to how she is, that she is human to what she is.

[6] Some readers may still inquire: Will not all primal truths be necessary for those who do acknowledge the meaningfulness of *de re* modal predications? If so, would not this show that the notion of primal truth has a modal underside after all? We should separate here two questions, a separation exemplified by the case of the analogous thesis, "All known propositions are true." First, there is a ques-

II.3 Individuation versus classification: a larger perspective. A modal metaphysician who assents to the necessity of Queen Elizabeth's being a caucasian but questions the necessity of her origin or assents to the necessity of this artifact's being a table while questioning the necessity of its being wooden, need not, most probably does not, question Saul Kripke's *notion* of necessity. He is simply more modest about its extension. Similarly here: the individuation and classification theorists share their notion of primal truth. To wit: a proposition '*x* is *P*' is a primal truth iff it has to be actually true that if *x* exists then *x* is *P*. What they differ about is the status of some such *P*s.

I shall not try to convince the reader that my more modest extension is the correct one. What is more, which extension one ultimately settles for may be secondary in importance to a proper understanding of the notion involved. The individuation and classification theorists share the larger picture: (i) they agree on the kind of question 'what is x?' is and on its autonomy *vis-à-vis* modal and epistemic matters, (ii) they agree on the pretheoretic intuition that propositions like, e.g., 'N. is human', 'Cayster is a river', and 'Whales are mammals' are true solely in virtue of what, not how, their subjects are. Pretheoretic agreement on such paradigm cases and on the basic notion involved—a trait *x* has actually to have, if *x* is actually to be—may be of more significance than differences over controversial cases motivated, in the first place, by theoretical desiderata.[7]

---

tion about the extensions of the relevant notions, viz., do all known propositions happen to be true? The answer is 'yes'. Quite a different question is about the two notions: Can sheer conceptual analysis of the subject notion, "known proposition," reveal its subsumption under the predicate notion, 'true proposition'? The answer is again 'yes'. In our case, things are rather different. As noted in connection to *de re* modal skeptics, conceptual analysis of the notion "primal truth" will not certify its subsumption under "necessary truth." If anything, it is by reflecting on our notion of necessity that one may argue that, when projecting necessities about *x*, we constrain them by truths about what *x* is. Far from exposing a modal underside in the notion of primal truth, the notion of necessity would be found to have a primal underside. Now these observations are all on the assumption that, at the level of extension, all primal truths are indeed necessary. Are they? According to the next section, logical truths are a special case of primal truths. In my "Logic and the World," *Journal of Philosophical Logic* (1989), contingent logical truths have been presented; hence there would be contingent primal truths. Are there then nonlogical contingent primal truths? I cannot address this matter here, involving as it does a consideration of propositions more complex than subject-predicate propositions and hence, a violation of the limits I set myself in this elementary exposition. I hope to return to this in future work on the very idea of necessity.

[7] (To the technically minded reader.) One additional way in which the classification and individuation theorists are brought here under a single roof is this: the

III. THE PRIMAL AND THE STRICTLY LOGICAL

III.1 Basic versus derivative primal truths. Much mental agility is called for tackling 'what is x?' questions. We do not want Nastassja and (the planet) Venus as same-whats just because, down the line, both are extended; nor do we want (the dog) Checkers and (the dolphin) Flipper as what-mates even though both are mammalian. On the other hand, most of us would not want to see Nastassja and Winnie (Mandela) segregated just because Nastassja is caucasian but Winnie is not; nor will we allow invidious distinctions in respect to whatness between Nastassja and St. Thomas just because she lives in the twentieth century and he lived a few hundred years earlier or because he was a "he" and she is a "she." We want Venus basically classified as a planet, Checkers as a dog, and St. Thomas, Nastassja, and Winnie as the same-whats: human beings.

A natural distinction between the *basic* and the *derivative* senses of 'what is x?' imposes itself. In its intended, basic sense, we aspire with 'what is x?' to the most minimal answer that is not too minimal. In contrast, when we respond with 'a mammal' for the last four mentioned items or with 'an extended thing' for all of them, it is merely a derivative sense of the question that is addressed. Yet again, the idea of such a distinction is acceptable to the individuation-oriented philosopher; he simply disagrees, in particular cases, on whether my candidate traits are basic enough.

Just as we distinguish basic and derivative senses of the question, I suggest to distinguish basic and derivative primal traits and truths. 'Nastassja is human' and 'Venus is a planet' are basic; there follows a

---

identity conditions of the latter are seen as a special case of existence conditions, viz., existence conditions that are uniquely specifying. This approach contrasts with the more standard view on which, if anything, existence conditions for *x* are seen as stated in terms of *x*'s identity conditions, viz., it is required that *x*'s identity conditions be "instantiated." On this last picture, the actual Nastassja (say, specified by 'offspring of gametes *N*') and her merely possible twin sister ('offspring of gametes *M*'), have been equally provided with identity conditions; their difference lies in that the former's, but not the latter's, identity conditions have been instantiated. This contrasts with the picture I have presented which insists that there is all the difference in the world between what actually is and what actually is not. For those that are, there are existence conditions and, if the individuation theorist is correct, in some cases such conditions might be uniquely specifying. On the other hand, that which is not is not. There is no "it" for which conditions, individuative or classifactory, can be stated. On my picture, "existence precedes identity," i.e., either *x* exists or not. If *x* exists, the truth that it does, "*x* exists," is basic and identity facts involving *x* presuppose it (to have *x* and *y* enter the identity relation, they must actually be). Thus, for *x* to satisfy 'to be identical to *y*', *x* has (i) to be and (ii) to be identical to *y*. On the other hand, if *x* does not actually exist, there are no truths involving "it"; in particular no truths about what "it" is identical with nor truths about what "it" is.

chain of truths, all primal, all stating conditions that have actually to obtain if their subjects are actually to exist, e.g., 'humans are mammals', 'mammals are animate', . . . and . . . '*X*s are extended'; or, 'planets are heavenly bodies', . . . and . . . '*Y*s are extended'. Consequently, 'Nastassja is extended' and 'Venus is extended' are primal truths but derivative at that. At the limit of such "derivations," we reach the strictly logical truths about Nastassja and Venus, e.g., each is, is self-identical, is red or not, is actually red if red, etc.

This last point relates to my introductory note and its record of an original hope to forge a connection between the primal and the "strictly" logical. On the traditional approach to the matter, such a project is discouraged from the very beginning. For we traditionally begin by investigating the strictly logical truths, attending to purely formal traits of Nastassja, traits she shares with all beings as such. Truths like 'Nastassja is human' are thus sentenced to remain out in the cold. On the present picture, we begin, classificationists and individuationists alike, with 'what basically is N.?'; so we also proceed with any other subject, Venus and Sirius, whales and snails. We then go on to more general traits shared by some of these subjects ('are animate'), to yet more general traits shared by all ('are extended'), eventually ending up with the strictly formal traits ('are self-identical'). All these propositions now naturally belong in the same kind, viz., predications that have to be actually true, if their subjects are to actually exist. The primal truths at the limit, the "strictly" logical truths, are seen to be "as concerned with the real world as truly as the truths of zoology, though with its more abstract and general features" (IMP, 169).

A pleasing, "technical" codification of the foregoing suggests itself. On the standard view, it is the idea of logical truth that is taken as basic. A primal truth is defined as the logical truth of a conditional 'If *A* then ϕ', with *A* some fundamental assumptions we hold fixed. On the present picture, the idea of logical truth emerges as the derivative notion, viz., ϕ is a primal truth closed under certain universal permutations of its individuals and attributes. This reversal of cart and horse carries some irony with it; and irony aside, it may hold the key for decoding the, at least a, thought behind "Logic is prior to how not to the what." First though the irony.

Logic, I here mean not so much what the mathematical logician practices but what the philosophical logician likes to theorize about, is often heralded as free of any *existential involvement*. By this last intended are both the object-language thesis that singular existential truths recording what there is, e.g., our 'Nastassja exists', are the epitome of merely material, hence nonlogical, facts; also, the metalanguage thesis that logical truths have no business committing us to the existence of anything. Even Russell, barely a few pages (glasses?) after his bold worldly *leitmotif,* recants. First, we are told that, though the primitive propositions of *Principia Mathematica* allow the inference that at least one individual exists, he now (1918) "regards it as a defect in logical purity." Harsher even, "there does not seem any logical necessity why in fact there should be any world at all" (IMP, 203). This is the gentle news, what with Russell at least trying to fly the worldly kite, then getting cold feet. More middle-of-the-road writers would simply dismiss as a charming joke the idea that logic is as concerned with the real world as zoology.

Charming or not, a joke it is not, at least not by the present lights. To the philosophical logician, existence, used or mentioned, may be the ghost in the logic ("inference") machine. Quite the contrary, says I: what is and what it takes for it to be are the very *terrain de chasse* of logic. It hunts the most fundamental, abstract and general, features of real things and the world they make up; so fundamental, that without them gone is the sense that those things *are,* that the world they make *is*. This focus on what is and what it takes for it to be shows in what I think of as the logician's, not the zoologist's, trap, the aforementioned frame: "It has to be actually true that: if *x* exists, *x* is *P*." As before, there is the metalanguage side of the matter with the trap designed to catch what is a *sine qua non* for Nastassja, by which 'non' is meant no *being,* absolute nothingness, for the old girl. There is also the object-language aspect with '(actually) exists' primed as a most obvious value for '*P*'.

I am now in a position to speculate on the connection between Nastassja's logical traits and what she is, a connection inspired by the mysterious yet seductive "logic is prior to the How, not to the What." As I see the connection, it is induced by two separate links—between what Nastassja is and existence conditions for her, a link discussed early in this work, and the just cited tie between her logical traits and, yet again, existence conditions for her.

Logic is seen to be prior to Nastassja's how not her what because, in studying what she is and her logical features, we are pursuing the same quest, viz., her *existence* conditions. It is simply that, in the logical case, the conditions are so abstract and general that they are universalizable. In both cases, we prowl prior to the how because we are focused on the primeval stage of the *provision* of subjects, not the eventual *description* of those subjects.

All in all, and lest there be any misunderstanding—my task is not to push the frontiers of logic so that they embrace Nastassja's humanity and other such treasures. 'Graceful-or-not' is logical and universal in extension, 'human' is not, and there is no question about this. The question is rather—should we canonize the extent of the extension of a predicate—is it universal or not—as our *quitte-ou-double* question? Elevating 'graceful-or-not' to a lordly position and consigning 'human' and 'featherless biped' to equal oblivion? Or should we rather start with the "meaning" of the predicate, I would prefer the property, involved, regarding all the while the extent of the predicate's extension as a corollary, flowing from how general and abstract a condition it is? I have been urging us this less trodden path.

IV. AN IDENTIFICATION PROCEDURE FOR PRIMAL TRUTH

IV.1 Cases. The notion of primal truth differs both qua notion and in extension from both apriority and necessity. Like primal truth, I believe that neither the notion of necessity nor that of apriority admit definitional analyses. What can and has been supplied for this pair of notions is a "computable criterion" for membership in their extension, viz., a method for "lassoing" the class of relevant truths. Such is the role of extensional surrogates, those which replace the "uncomputable" intensional original by a "computable" extensional surrogate, viz., " 'x is P' is necessary iff 'x is P' is true in all possible worlds" and " 'x is P' is apriori iff 'x is P' is true in any evidential situation (context)." Can we provide a similar, "extensional" and "computable," identification method for primal truth?

Consider the question: What do we *do* when faced with a claim that x has actually to be P if it is actually to exist, or, less theoretically, that what x is is P? I suggested that we engage in a counteractual subtraction of P from x's profile and see whether it is stopped in its tracks. But what is a counteractual subtraction? *Where* do we separate x and P when we so subtract? After all, in the actual, real world, x is P. The answer is: we separate them in a *story*, in our local vernacular, in a *case*, one that has a P-less x as its leading character. Systematic failure to come up with such a countercase will lead us to suspect that there is truth in the what-claim. As I read our practices, they suggest that:

(P) 'x is P' is primal iff all coherent cases about x make it P.

IV.2. Hard Cases. (P) is meant to settle what-conjectures by running through cases. Inevitably, some of the cases will turn out to be hard cases. To make no bones about the problem, consider what, in my eyes and to my knowledge, is probably the hardest case of all, viz., Descartes's question 'what am I?' and his subsequent story about himself, there by the fire, but his body, his very extendedness

I should say, shed away. A story he took to show, at the very least, that being extended is not part of what he is (and not merely, as the modalists instruct us, that it is not part of his "modal essence"). A coherent case? A hard case above everything else. Alas, some of us, perhaps with the ideal of a *Methodenreinheit* in mind, tend to leap from the existence of such hard cases to skepticism about the very method.

My apologia is that the purity, let alone power, of a philosophical method should not be evaluated with the false ideal of a universal-decision procedure in mind. Intrinsically, even if seemingly elementary, hard cases are just the natural breaking points of reliable methods. We do not dismiss our computing methods when a diophantine equation in 9 variables turns out to elude them. We do not dismiss the method because we begin to suspect that there may be no better one. This, in turn, forces us to reflect on the reasons that made this particular school-boy arithmetic problem recursively unsolvable; we then reflect some more on the nature of the method itself and come to see why it had to break down at this point: if it did not, something too good to be true would have been happening. We mature. We realize what the method can do and what no method can undo.

What is good for the number theorist is good for the philosopher: Descartes's case should come out a hard case, perhaps intrinsically unsolvable. Were the method to suggest that things here are crystal clear rather than as seen through a fog, I at least would be instantly suspicious of the method. Far from being an embarrassment, hard cases (of the Cartesian delicacy) provide the method with the seal of reliability. Instead of lamenting the existence of a few hard cases, we can bask in the verdicts delivered by most of them. Using his own favorite method, we can show that it is not part of what Descartes is to be a philosopher or/and a mathematician; nor are, and this is more intriguing a conclusion, the modally essential traits of being French, being Caucasian, living in the seventeenth century, and originating in egg D. In contrast, to be a thinking being and a human being are traits that are inseparable from him, so most of us feel, and thus should be counted as primal. The cases involving being in time, being extended, mammalian, etc., might be left unsettled. Not unsettled for me; but I should like it noted that we can give the method a clean bill of health even on the assumption that these cases are left undecided.[8]

---

[8] As hinted, here and below when the case of a painless c-fibers' firing is discussed, I believe that the present framework may enlighten us about the mind/

IV.3 Countermodels and counterexample cases. The method of cases is to the philosopher what the method of (counter) models is to the mathematician and the logician. All aim to perceive, clearly and distinctly, connections between propositions, in particular, whether a target proposition *S* is already "contained," or, in contrast, "separable from," a body of assumed propositions *B*. In investigating "containment" and "separability," the logician abstains from any modal involvement. Thus, the axiom of choice is *modally* dependent on the base axioms of set theory: in any world where the latter hold, it holds, too. Nonetheless, it is *logically* independent from those axioms, there being models verifying the base axioms but falsifying it.

Cases, the philosopher's models, display the same coupling of modal dependence with (onto)-logical independence: 'Descartes is a seventeenth-century figure' is modally dependent on 'Descartes is human'. Nonetheless, the former is ontologically independent of the latter, as a coherent story about that man living a century earlier or later testifies. All the cases fabricated in our search for what Descartes is may turn out to be genuine modal possibilities for him; nonetheless, it is not part of the point of spinning such cases to engage in the metaphysics of modality.

Cases can and do transcend "the modal horizon." To ground the point, it may be worthwhile to reflect on some classical cases. Suppose it turned out that the story about twin earth represented no genuine possibility for our universe. Perhaps because of the modal determinism of footnote 5, i.e., given that things could not have been otherwise and that there is no twin earth, there could not have been one; or as I prefer, simply because there is no point in the history of our universe where it could have gone on to produce an exact replica of our (already vastly improbable) earthly development. With twin earth representing no genuine possibility, would the point of Hilary Putnam's tale be gone by the board? Not as I read it. The point of the case is, to begin with, ontological, not modal, viz., for there to be water in actuality is one thing, any old liquid with waterish appearance quite another thing. This much is a

---

body distinction. I can even be more specific while remaining mysterious: the what-I-am versus how-I-am distinction affords a reconstruction of (one of) Descartes's arguments for the real distinction, a reconstruction that may well vindicate the intended distinction. At least so I argue in "Philosophical Fiction and Real Distinctions," the aforementioned work dedicated to such a reconstruction of Descartes. I hope to convince myself, sometime in the next few years, of the conclusions of my reconstruction.

premodal observation on what it takes for water actually to exist (its not sufficient that a waterish liquid shows up); this is not a modal claim about how our world, with the water it does actually have, might have been.

Like remarks apply to other classical cases; e.g., post-Kripkean critics have charged that a case he has presented, of Jones's c-fibers firing but no pain felt, may have no bite because a necessary connection, of course only empirically discoverable, may tie the firing of the fibers and the feeling of pain. In response, I say: let there be such a necessary connection; it does not rob Kripke of his ontological point, whether or not he himself intended to read the situation in this modal-free way. A coherent story can be concocted, it seems, about Jones's agitated fibers but no pain felt. This much has been conceded by physicalists who, in the old, pre-Kripke, days, agreed that there is nothing "conceptually incoherent" in a painless fibers' firing. I "translate" and concur: by 'conceptually incoherent', we mean neither "is a priori excluded" nor "is modally impossible." We mean rather "is barred by *what* the items are." Then indeed there is nothing in what c-fibers' firing is to bar a painless firing. In contrast, there is much to bar a firing that involves no atoms, is not located, consumes no energy, etc.

And so it goes: we are told not of one person but two, waking up two hundred years after the death of Guy Fawkes with memories (qualitatively) identical to Fawkes's; we are privy to Einstein's view of the universe as he is riding a light ray or the life and times of us, humans, in a two-dimensional space. With my guardedness over the span of the possible, I see in none of these stories a way the things themselves might have been. No matter. The point of these cases is not to speculate on the merely possible ways the world might have followed but to divulge secrets about the fabric of things here in actuality.[9]

V. A CONCLUDING NOTE

Cases provide us with "graphic evidence." They are the philosopher's brand of test tube, his way of probing the elusive chemistry of subjects like Nastassja. I remind the reader it is not the existence of a case, let alone its discovery, that *makes* a truth (non) primal. The critical question is whether P is an existence condition for x, i.e., whether it has to be actually true that, if x exists, x is P. At an even more primordial level, I would regard my own elucidation as somewhat derivative. At that ultimate level, we would find that we had come full circle back to the opening page of this work, where intuitions about what, as opposed to how, Nastassja is were alluded to. At this level, what makes an arbitrary proposition of the form 'Nastassja is *P*' true solely in virtue of what she is is what she is.

---

[9] Many, to judge by current metaphysical discussions, would allow that the cases represent no "metaphysical possibilities" but nonetheless are "logical possibili-

JOSEPH ALMOG

University of California/Los Angeles

---

ties." I confess that I am rather alienated by this popular mode of expression. In this paper, as elsewhere, I mean by 'necessary' "absolutely necessary," what, I think, Kripke means by his insistence on "necessity tout court." In particular, to recall the above example, the axiom of choice is necessary tout court (this is meant to exclude the Fraenkel-Mostowski countermodels, not just the later, pure sets, countermodels, as possibilities). To claim that the axiom of choice could have been, "logically" as it were, false, is, in my eyes, to confuse the metalanguage, nonmodal, claim, that there is a model of set theory, where it is false, with the object-level claim, concerning ways the world, even if abstract facets of it, might have been. Like remarks apply to the induction axiom and the existence of nonstandard models of arithmetic where it goes haywire. Formal independence proofs are simply no proofs of genuine possibility. Closer to home, recall that the axioms of propositional normal modal logics, as strong as S5, allow one-world models, models used in relative consistency proofs for such systems. But assume, as is only natural, that the metaphysics of our (fn. 5) modal determinist is false, that things could have been otherwise. Then by none other than the intended meaning of the S5 axioms, the one-world setting is not a genuine possibility. In my view then, much of the currently free talk of "logical" possibilities stems, ironically, from mistaking use and mention, the original sin in this land of confusion: consistency talk, which is (i) nonmodal talk about (ii) sentences (theories), is blended with (i') modal talk about how (ii') things, concrete or abstract, might have been. As I see it, models are models and possibilities are possibilities, the two simply belong in different strata. The idea that models are not modal and that, by simple independence results, there are many necessary (mathematical) truths for which countermodels are constructible is developed in detail in "Philosophical Fiction and Real Distinctions."

'P, THEREFORE, P' WITHOUT CIRCULARITY*

My main thesis is that a variety of arguments having the form '*P*, therefore, *P*' do not beg the question. A consequence is that syntactic theories of circularity are false and pragmatic ones are confirmed. A third theme of this paper is that some of these arguments owe their surprising persuasiveness to a neglected feature of arguments: exemplification.

I. RIVAL THEORIES OF CIRCULARITY

The simplest definition of 'circular argument' is "argument whose conclusion is also a premise." Proponents of the syntactic approach like the spirit of this definition because it portrays circularity as a relationship between the components of the argument. They favor analyzing circularity as a formal fallacy, such as affirming the consequent or undistributed middle; syntactic theorists say that whatever is bad about circular arguments is due to the *form* of the argument.

Despite their sympathy with the basic idea of the simple definition, syntactic theorists agree that it is too narrow. The definition only states a sufficient condition because it fastens on the most salient cases of circularity. Wild specimens of the fallacy conceal the repetition. Usually the identity of premise and conclusion is disguised by reclothing the conclusion in different terminology or grammar. Sometimes the conclusion is smuggled into the premise set by disassembling it into unfamiliar looking parts. Judicious juxtaposition with other premises also aids the illusion.

These complications are met with amendments aimed at outlawing these subterfuges.[1] This is as tricky as writing loopholes out of a tax code. To avoid condemning most valid deductions as circular, you must be specific. But it always seems to be just a matter of time before someone weasels through the clauses. When you then interpret the rules loosely to achieve a *de facto* elimination of loopholes, critics complain that you are exploiting the vagaries of propositional identity in an *ad hoc* fashion.

---

* An ancestor of this paper was read at the Eastern Division APA meeting in 1989. It has been improved by the reply of my commentator, Virginia Klenk. I also wish to thank Jonathan Adler, John Carroll, Julia Driver, David Sanford, Douglas Stalker, Kent Wilson, Allen Wood, and the editors of this journal for their comments and suggestions.

[1] For a recent example of a syntactic definition of 'circular argument', see p. 49 of Kent Wilson's "Circular Arguments," *Metaphilosophy*, XIX, 1 (January 1988): 38–51. Since Wilson distinguishes between circularity and vicious circularity, he would not regard himself as legislating against circularity. He does regard question-begging as always fallacious and so would reject my belief that 'circular' and 'question-begging' are synonymous.

0022-362X/91/8805/245–66

This mess led many syntactic theorists to develop dialogue games. The strategy is first to figure out which rules will preclude circularity in simple, formal dialogues and then to regulate progressively enriched language games.[2] The dialogue format introduces greater scope because disputation often involves relationships *between* whole arguments. Codification of rules and moves controls the complexity costs of the enhanced realism. The gradualism behind moving from simple to enriched games also strikes a nice compromise between piecemeal and systematic attacks on problems.

The pragmatic approach views circularity as a relationship between the argument and the people using it. For example, David Sanford[3] says that

> [Question-begging] is not a purely formal matter. An argument formulated for Smith's benefit, whether by Smith himself or by another, begs the question either if Smith believes one of the premises only because he already believes the conclusion or if Smith would believe one of the premises only if he already believed the conclusion. A primary purpose of argument is to increase the degree of reasonable confidence which one has in the truth of the conclusion. Not every argument which fails this purpose begs the question, but every question-begging argument fails this purpose (*ibid.*, p. 198).

This passage contains two sufficient conditions for circularity and a necessary condition, but not a condition that is both necessary and sufficient. The conditions featured in this partial analysis are broad, however, and accord well with views of other writers who have emphasized the epistemic aspects of inference.[4] Other pragmatists provide more detailed accounts of how circular arguments fail to be knowledge-expanding. William Lycan and George Schumm[5] say the problem is that the conclusion only differs trivially from the premises: circular arguments are valid but uninformative. They say that the vice of circularity is dialectic inefficiency. Frank Jackson[6] analyzes circularity as egocentric reasoning: the arguer selects premises in accordance with his own background beliefs rather than his audience's. Jackson also analyzes many apparently circular arguments as instances of "misleading advertising" because they have argumentative implicatures that mislead the audience about the nature of the arguer's evidence.

---

[2] C. L. Hamblin introduced the dialogue approach in *Fallacies* (London: Methuen, 1970). J. D. Mackenzie has refined the account in a series of articles, including "Why Do We Number Theorems," *Australasian Journal of Philosophy*, LVIII, 2 (June 1980): 135–49. Also see Frank Van Dun, "On the Modes of Opposition in the Formal Dialogues of P. Lorenzen," *Logique et Analyse*, LVII–LVIII (1972): 103–36; and Lauri Carlson's *Dialogue Games* (Dordrecht: Reidel, 1983).

[3] "Begging the Question," *Analysis*, XXXII, 6 (June 1972): 197–9.

[4] In particular, Sanford cites W. E. Johnson's *Logic*, Susan Stebbing's *A Modern Introduction to Logic*, and Max Black's "Induction and Experience," in *Experience and Theory*, Lawrence Foster and J. W. Swanson, eds. (Amherst: Massachusetts UP, 1970).

[5] "A Theory of Question-Begging" (unpublished manuscript).

[6] *Conditionals* (New York: Blackwell, 1987), ch. 6.

Pragmatists do not agree with the spirit of the simple definition of circularity. Although they grant that the definition provides a simple sufficient condition for circularity, they lament the success of this condition on the grounds that it misleads people into thinking that circularity is a formal fallacy. Pragmatists think that the syntactic theorists have been mesmerized by the cases involving blatent repetition of the conclusion. The one-sided diet of examples fed to us by the sufficient condition encourages us to assimilate all cases of circularity to these unrepresentative "paradigms" of circularity. Compare this misguided reduction to our tendency to view all words as names. Pragmatists deny that any amount of tinkering with the simple definition will yield an adequate definition. They insist that charges of circularity must be relativized to audiences with varying purposes, background beliefs, and inferential prowess.

I am a gung-ho pragmatist. I think my brethren are too soft on the syntactic theorists. We should totally reject the simple definition that starts the syntactic ball rolling because propositional repetition is not a sufficient condition for circularity. Indeed, we should hit them where it really hurts. We should deny that all '*P*, therefore, *P*' arguments are circular. If the syntactic theorist cannot assume that '*P*, therefore, *P*' suffices for circularity, he has to close up shop; for this is the simplest case of all, the seed value for all his formulas.[7]

II. EXAMPLES THAT DO NOT WORK

A genuine counterexample to "All '*P*, therefore, *P*' arguments are circular" must express the same proposition in the premise and conclusion. We can achieve orthographic identity through devices of equivocation, such as polysemy (multiple senses), homonymy, and indexicality. But orthographic identity is not enough for propositional identity. Assume that in (A1) 'child' is ambiguous between 'offspring' and 'immature offspring' (respectively), that in (A2) 'ring' is homonymous between sounds and jewelry, and that in (A3) 'this' is self-referring.

---

[7] John Woods and Douglas Walton advocate an eclectic approach in "Arresting Circles in Formal Dialogues," *Journal of Philosophical Logic*, VII, 1 (February 1978): 73–90. The dialogue approach cannot retain its identity and accept my claims about the noncircularity of some '*P*, therefore, *P*' arguments. The rules would have to become highly pragmatic so that the resulting system would look more like a web of Gricean maxims than a formal game.

(A1) Steven is Lori's child.
Steven is Lori's child.

(A2) John recalled two rings.
John recalled two rings.

(A3) This is the first token of this sentence.
This is the first token of this sentence.

These readings make all three of these arguments noncircular (and indeed invalid). They do not really have the '*P*, therefore, *P*' form, however.

III. EXAMPLES THAT WORK

I now turn to examples that support my thesis that some '*P*, therefore, *P*' arguments are not circular. To indicate my judgment of their relative plausibility, I begin with the strongest and work down to the weakest.

A. Self-supporting versions. Some arguments are self-supporting in the sense that features other than validity buttress the conclusion:

(B1) If there are syllogisms, then information can be combined.
There are syllogisms.
Information can be combined.

(B2) An argument with a missing premise is an enthymeme.
There is an enthymeme on this page.

Since (B1) is an argument with two premises, it is a syllogism and hence evidence for its second premise, which in turn enhances the credibility of the conclusion. Argument (B2) is itself an enthymeme and so ensures the truth of its own conclusion. The following '*P*, therefore, *P*' argument is self-supportive in the same way:

(C1) Some arguments are written in black ink.
Some arguments are written in black ink.

In addition to be being sound, (C1) is rationally persuasive. A person who doubted the conclusion comes to know it is true by considering the argument. Since no rationally persuasive argument is circular (recall Sanford's necessary condition), (C1) is a noncircular '*P*, therefore, *P*' argument.

Self-support varies in kind and degree. Witness the mild self-help of

(C2) MORE THAN ONE ARGUMENT IS WRITTEN IN CAPITAL LETTERS.
MORE THAN ONE ARGUMENT IS WRITTEN IN CAPITAL LETTERS.

This argument token helps itself by filling half of its own quota but it does not fully confirm itself. Another parameter is the universality of self-support. We can capture this notion by defining self-support for argument types in terms of self-support for argument tokens. Call an argument *type* universally self-supporting exactly if all its tokens support themselves. Say that the type has limited self-support just in case only some of its tokens are self-supporting. The (C1) and (C2) argument types have limited self-support because there are green tokens of the (C1) type and "de-capitalized" tokens of the (C2) type. Examples of universal self-support can be obtained by exploiting formal properties of arguments:

(C3) Some arguments are composed solely of existential generalizations.

Some arguments are composed solely of existential generalizations.

(C4) It is possible for a conclusion to entail its premise.

It is possible for a conclusion to entail its premise.

(C5) Some deductive arguments do not reason from general to particular.

Some deductive arguments do not reason from general to particular.

Since (C3)–(C5) owe their persuasiveness to their logical form, they should be acceptable to even those with a narrow conception of logic. Arguments exploiting pragmatic properties make up an intermediate case:

(C6) Someone has thought about an argument.

Someone has thought about an argument.

(C7) An argument has been propounded by someone at some time.

An argument has been propounded by someone at some time.

Anyone who thinks about (C6) thereby confirms it. Since all tokens of the (C6) type are confirmed in this obvious fashion, it is a self-supporting argument type. If you propound (C7), it is also rationally persuasive. But if a (C7) token is merely mentioned, as when we consider an argument as an abstract entity, then it is not self-supporting. But since we tend to consider arguments as if they had been propounded, (C7) is apt to strike us as a self-supportive argument type.

Self-supporters can also be driven by properties of their *components:*

(D) Either no one can read or some people refuse to follow directions.
Someone can read.
Some people refuse to follow directions.

Reading the second premise confirms it and hence strengthens the argument. Now consider a '*P*, therefore, *P*' version:

(E1) Some sentences are exactly seven words long.
Some sentences are exactly seven words long.

In (E1), the real persuasive work is done by a property of the premise alone rather than by the argument. But when the statements are relationally self-supporting, both components play a role:

(E2) There are at least two tokens of an eleven-word sentence.
There are at least two tokens of an eleven-word sentence.

The premise and conclusion work together in (E2). Nevertheless, they would be just as persuasive if they had been merely listed. Since many people are more impressed by examples where the rational persuasion is done by an argument in its role as argument, they prefer the earlier examples.

Self-supporting '*P*, therefore, *P*' arguments are seldom propounded because we are rarely interested in establishing the sort of conclusions they demonstrate. But there are a few used in ordinary life. For example, I challenge your claim that clocks are to be set forward in spring and back in fall (as part of "daylight savings time"). You reply, "Spring forward, fall back."[8] The proposition expressed as your reason is identical to the conclusion. Nevertheless, the argument is persuasive because your premise is couched in a style typical of mnemonic sayings. My best explanation of your premise is that you are manifesting memory. For a second case, suppose you claim that Ronald Reagan phoned you last night. I confess that it was a practical joke: "It was only I imitating Ronald Reagan." You demand proof. I reply by repeating my confession in my Ronald Reagan voice. Similarly, I back 'I speak French' with 'Je parles francais'.

---

[8] J. D. Mackenzie reports persuading a challenger that April has only thirty days by replying, "Thirty days hath September, April, June, and November"—"Why Do We Number Theorems?" *Australasian Journal of Philosophy*, LVIII, 2 (June 1980): 135–49; p. 140. The form of this argument is '*P* & *Q* & *R* & *S*, therefore, *Q*' rather than '*P*, therefore, *P*', so I had to use a slight variant.

Grumps are leery of the (C) and (E) examples on the grounds that they involve self-reference. But this is 1991! 'Self-reference' is no longer a dirty word. Contemporary commentators on the liar paradox agree that Russellian knee-jerk rejection of self-reference would deprive us of substantial theoretical and practical benefits. Nevertheless, philosophers who view the liar paradox from a distance latch onto the self-referential element and do not let go. They know that *something* is deeply awry with 'This statement is false', they know that the liar statement is self-referential, and so come to have a vague but tenacious distrust of anything that smells of self-reference.

The assertions in the (C) arguments refer to properties of *arguments,* not assertions, and so are not about themselves. In the (E) cases, the assertions are about properties of assertions but do not refer to themselves with *singular terms*. The conclusions of the (E) (and (C)) arguments are existential generalizations and, although such statements are made true by individuals, they do not refer to them; 'Some men are divorced' does not refer to Johnny Carson. Sure, the (C) and (E) arguments do refer to themselves in the stretched sense of containing *predicates* whose extensions include the arguments or constituents of the argument. In this wide sense, 'There are visible things' is self-referential because it is visible. But once 'self-reference' is construed this broadly, it can hardly be a diagnostic term that picks out a property warranting suspicion.

The griping is really set off by simultaneous exemplification and denotation.[9] Each (C) and (E) argument (or its constituents) exemplifies the predicate its conclusion claims to be instantiated. Exemplification is narrower than the weak sort of "self-reference" discussed in the previous paragraph because exemplification is selective. Unlike the (C) and (E) arguments, the following does not exemplify the predicate asserted to be instantiated by the conclusion.

(F) Some arguments are typed left-handed.
Some arguments are typed left-handed.

In this case, the problem is inaccessibility: you cannot readily tell whether (F) instantiates 'typed left-handed', so (F) is not an exemplar. In contrast, the capitalization of (C2) is ostentatious.

---

[9] The relevance of Nelson Goodman's notion of exemplification was brought to my attention by Douglas Stalker. As my talk of propositions suggests, I have detached exemplification from Goodman's nominalism. For an extended discussion of exemplification, see Catherine Elgin's *With Reference to Reference* (Indianapolis: Hackett, 1983), ch. 5.

Exemplification is a subconverse of denotation: a denoting predicate refers to its object, while an exemplar is an object that refers to its predicate *and* instantiates it. So model homes, promotional handouts, and sample proofs all qualify as exemplars. Philosophically richer examples include Thomas Kuhn's paradigms, Iris Murdoch's moral exemplars, and the aesthetic works analyzed by Nelson Goodman. Some words simultaneously denote and exemplify, consider 'word', 'short', 'polysyllabic', and onomatopoeia such as 'bow-wow' and 'ding-dong'. The (C) and (E) arguments also appear to denote and exemplify simultaneously. For example, (C1) exemplifies the predicate 'argument written in black ink' used in the conclusion, while this predicate also denotes the argument. Normally, the interesting properties of an argument are what its premise and conclusion say. The (C) and (E) arguments show that the predicates exemplified can also be important. There is a parallel with poetry. It is natural to concentrate on what a poem says and to overlook what it exemplifies. As Goodman[10] observes, translation of poetry sometimes brings the omission to attention:

> The distinction illustrated here [between samples and labels] becomes important in the translation of poetry or other literature. Original and translation will of course differ in some properties; but so do any two inscriptions of the same word, or even any two red inscriptions of "red". The goal is maximal preservation of what the original *exemplifies* as well as of what it says. Translation of a staccato by a legato text may do greater violence than some discrepancies in denotation (*ibid.*, p. 60).

The point can be extended to argument translation. The natural Japanese translation of (E1) contains only six words, not seven. So, if this article were translated into Japanese, the sentence's description of sentence length should be revised to six words. Or suppose I wish to lecture on (C1) using a blackboard: 'white chalk' should be substituted for 'black ink'.

Self-referential worries are not the only obstacle to acceptance of the (C) and (E) cases as counterexamples to "All '*P*, therefore, *P*' arguments beg the question." One vague suspicion is that an argument such as (C1) *causes* belief in the conclusion without providing a *reason* for the conclusion. One can imagine a superneurologist causing you to believe that giraffes swim with the carefully calculated optical properties of the inscription 'Toads smile'. But his 'Toads smile, therefore, giraffes swim' argument would not give you reason for the conclusion even though it triggers belief in it. Although I agree that this story shows that there is a difference between merely causing belief and giving reason to believe, I deny that the distinction cuts against (C1) or any of the other examples. After all, they do *rationally* persuade. The cause/reason distinction is a red herring.

---

[10] *Languages of Art* (Indianapolis: Hackett, 1976), 2nd ed.

Appreciating the irrelevance of this distinction may lead one to draw another distinction that is relevant. ‘Reason’ sometimes refers to a justifying proposition and sometimes refers to things the proposition is about. Thus I may explain my hospitalization with the propositional reason, ‘My leg broke’, or by pointing to my broken leg as my “ontic” reason. We are used to thinking of arguments as aiming to have premises that are propositional reasons for their conclusions, but the (B)-(E) cases show that an argument itself can be an ontic reason for its conclusion. The population of actually uttered arguments falls into four categories:

	Ontically persuasive	Ontically unpersuasive
Propositionally persuasive	Few arguments	Most arguments
Propositionally unpersuasive	Very few arguments	Many arguments

An argument that is ontically but not propositionally persuasive will strike some as a “conflict case” between two conceptions of argumentation:

(G1) The moon is edible.
———————————————
Some argument is written in black ink.

An objector might insist that we must *choose*. He will say that, for *logical* purposes at least, we should only consider propositional persuasiveness. So the objector concludes that the (C) and (E) arguments are circular after all: they rationally persuade in the wrong way (ontically).

Aversion to ontic persuasion is a symptom of validity chauvinism. If we read in a goal of perfect truth preservation, then (G1) is a bad argument in the sense that it fails to achieve the propounder’s objective. But imputing this goal to inductions equally well condemns them as bad arguments. As in the case of induction, charity dictates the attribution of another goal when evaluating (G1). In particular, the propounder of (G1) can legitimately aim at persuasion by exemplification. Ditto for:

(G2) Some arguments are written in black ink.
Some arguments are written in English.
Some arguments are self-supporting in two ways.

(G3) If this argument is black, then it is the color of coal.
This argument is the color of coal.

(G4) Marilyn Monroe had six toes on her left foot.
Some arguments have a premise that is superfluous to its validity.
Some arguments have a premise that is superfluous to its validity.

Like (G1), (G2) and (G3) are rationally persuasive but invalid. (G4) is valid but this feature takes a back seat to the aim of persuasion by exemplification.

The validity chauvinist might grant that an arguer can have ontic persuasion as a goal and yet complain that this goal is an *illegitimate* aim. He might buttress this charge of logical perversion with an appeal to the frequency with which people balk at arguments such as (G1). But notice that students in logic class also complain of arguments that persuade in the wrong way. They balk at valid arguments corresponding to the paradoxes of strict implication. Students hunger for relevance and so resist sound arguments having unrelated tautologies as premise and conclusion. They want efficiency and so are ambivalent about the result of adding even "relevant" premises to a deduction previously agreed to be acceptable. Students demand credible premises and so must be dragged through their first reductios. These misguided balks should make us wary of resting on the balk against (G1). Granted, (G1) is invalid and has a premise that is both false and not topically related to the conclusion. But so what? It does rationally persuade in its peculiar way and logic is about rational persuasion through argument. Should logicians not explore this path rather than zone it off as outside the domain of logic? The burden of proof is on the balkers.

I see only two plausible ways for them to shoulder this burden. First, insist that (C) and (E) arguments are all *enthymemes;* and say the missing premises in each are self-referential propositions to the effect that the argument (or its components) instantiates the predicate the conclusion claims to be instantiated. But this proposal violates the maxim telling us to add premises only out of charity. The (C) and (E) arguments are already valid and the extra premises do not free them from the charge of circularity. Indeed, filling in missing premises makes some of them look worse:

(C3′) This argument is composed solely of existential generalizations.
Some arguments are composed solely of existential generalizations.
Some arguments are composed solely of existential generalizations.

Whereas the original (C3) is sound and quite persuasive, (C3′) is unsound and less persuasive. Logician, do no harm!

The next objection appeals to meta-arguments rather than enthymemes: the (C) and (E) cases are instances of *indirect* rational persuasion. Thus (C1) only persuades us by bringing a meta-argument to our attention:

(C1.1) (C1) is an argument written in black ink.
Some arguments are written in black ink.

Notice that (C1.1) does not have the '*P*, therefore, *P*' form but does have the same conclusion as (C1). The allegation is that, since (C1.1) is doing the ultimate work of rational persuasion, (C1) is really circular after all. (C1) presents an illusion of cogency because it resembles an argument that is cogent and that is immediately brought to mind by (C1). The resemblance and immediacy make it seem as if only (C1) is involved in the persuasion process when in fact there are two arguments.

My first reply is that the objection assumes that indirect persuasion is not persuasion. Compare indirect persuasion with indirect perception. When I see the wind blowing through the trees, I see the wind even though I am only seeing it indirectly through its effects.

My second reply is a charge of special pleading: to postulate a meta-argument as the ultimate persuader for (C1) is selectively to enforce a superficially plausible principle of justification. The basic version of the principle is that a conclusion is justified only if its premise and the link between the premise and the conclusion is justified by another argument. Conscientious application of the principle would not stop with the postulation of (C1.1). Other meta-arguments would have to be marshalled to justify the premise of (C1.1), for example

(C1.11) (C1) is in black print.
Most black print is composed of ink (rather than laser burn).
(C1) is written in black ink.

After supplying arguments for the truth of (C1.1)'s premise, we would be obliged to supply arguments to establish its validity:

(C1.12) (C1.1) has the form of existential generalization.
Existential generalization is a valid argument form.
(C1.1) is a valid argument.

Now, at least, it seems we have satisfied the obligations imposed by the justification principle. One reward of our circumspection is that we see that even (C1.1) is only indirectly persuasive; for it depends on a cluster of background arguments needed to assure us that its premise is true and that the conclusion is truly implied. But on reflection, we see that our justificatory work has only begun! What about the soundness of the arguments supporting (C1.1)? Unless we accept them on blind faith, we will have to inquire into their logical credentials. And as long as we are going to do the job, let us do it right and inquire into subtler properties of arguments. We need to postulate further meta-arguments to reflect factors, such as the arguer's intentions, the conventions of disputation, and the argumentative implicatures conveyed by the selection and arrangement of premises and indicator words, such as 'since' and 'therefore'. Thus the metaargument objection is undermined by an infinite regress akin to Lewis Carroll's.[11]

B. Self-defeating versions. Whereas self-supporting arguments help themselves, self-defeating ones hurt themselves.

(H1) All arguments are written in green ink.
All arguments are written in green ink.

No circular argument is rationally dispersuasive. The complaint against circularity is that you are standing in place when you think you are going forward; not that you are going backward. Hence (H1) is a second example of '*P*, therefore, *P*' without circularity. Like self-supporters, self-defeaters can be affected by their formal properties:

(H2) No argument has a negative premise.
No argument has a negative premise.

(H3) All arguments are syllogisms.
All arguments are syllogisms.

Unlike (H1), both of these arguments are universally self-defeating.

The idea that you can weaken your case merely by arguing for it (or arguing for it in a particular way) is not esoteric. When Joe Sixpack is told that he is too defensive, contentious, or sophistical, he will worry that a rebuttal will confirm the criticism. If you tell Joe that he overuses the charge of circularity and you back up the accusation with the claim that he uses 'That's question-begging' as an argumentative crutch, he will be reluctant to object that your argument is circular. A variety of philosophical arguments have been alleged to be self-defeating. For each of the following, there are critics who say the position is undermined by the very act of arguing for it: behaviorism, cognitive pluralism, determinism, dogmatism, eliminative materialism, epiphenomenalism, epistemological anarchism, evolutionary epistemology, mechanism, nominalism, psychologism, solipsism. Ditto for arguments against free speech, objective truth, the meaningfulness philosophical discourse, and laws of logic. The charge of argumentative self-defeat is illustrated by Israel Scheffler's[12] critique of those who argue for the subjectivity of science on scientific grounds:

---

[11] "What the Tortoise Said to Achilles," *Mind*, IV, 14 (April 1895): 278–80.

> . . . there is a striking self-contradictoriness in the effort to persuade others by argument that communication, and hence argument is impossible; in appeal to the facts about observation in order to deny that commonly observable facts exist, in arguing from the hard realities of the history of science to the conclusion that reality is not discovered but made by the scientist. To accept these claims is to deny all force to the arguments brought forward for them (*ibid.*, p. 44).

Avoiding self-defeat is a *standard problem* for radical philosophical positions. This is evidence that appeals to self-defeat are part of accepted practice in argument evaluation and hence are of interest to logicians in the same way as appeals to validity.

No doubt, validity is far better understood than self-defeat. Much of the obscurity clears with careful definition and patient distinctions, however. Count an act as self-defeating just if it frustrates the achievement of its purposes. (Purposes are assigned to acts in the way they are to artifacts, such as pans, pens, and pins.) Thus it is self-defeating to struggle in quicksand and to use highbeams on foggy roads. Since an utterance is an act, we can infer that an utterance is self-defeating exactly when it frustrates the achievement of one of its purposes. Examples of linguistic backfire include:

1. The successor of the largest integer ever thought of is 192,394,031,331.
2. The fact that turtles gargle is ineffable.
3. It is possible that zebras hibernate but no one can conceive it.
4. Be spontaneous.
5. Never say 'never'.

---

[12] *Science and Subjectivity* (Indianapolis: Bobbs-Merrill, 1967).

Correct application of 'self-defeat' requires knowledge of what the speaker was trying to do and what he has actually done in the attempt. Since the speaker has many purposes and his utterance instantiates many descriptions, 'self-defeat' should be relativized to these purposes and descriptions. The consequent relativity of self-defeat is illustrated by 'I never warn' which is self-defeating *as a warning* but not *as an assertion*. Utterance descriptions grow in complexity. Slowly saying 'I never speak slowly' is self-defeating, as is slowly saying in a Texas drawl 'I never talk in a slow Texas drawl', and slowly saying in a Texas drawl on Sunday 'I never talk in a slow Texas drawl on Sundays'. The larger utterances are not self-defeating, however, relative to the shorter utterance descriptions.

The heterogeneity of speaker goals can be gleaned from the variety of speech acts. 'Never take advice' is self-defeating because it cannot achieve the standard point of advising: being fruitfully heeded. A principal purpose of asking a question is to learn the answer from the hearer. This makes 'Can anything be learned by asking questions?' self-defeating. The very act of asking presupposes belief in an affirmative answer and hence precludes the possibility of achieving the goal of learning the answer. The head quote of Irving Copi's logic text contains an excerpt from the *Discourses of Epictetus* featuring a request that is self-defeating for much the same reason:

> When one of his audience said, "Convince me that logic is useful," he said,
>
> "Would you have me demonstrate it?"
>
> "Yes."
>
> "Well, then, must I not use demonstrative argument?"
>
> And, when the other agreed, he said, "How then shall you know if I impose upon you?" And when the man had no answer, he said, "You see how you yourself admit that logic is necessary, if without it you are not even able to learn this much—whether it is necessary or not."

Many speech acts (assertion, warning, promise) strive for plausibility, so a wide group of utterances can be *evidentially* self-defeating. The degree to which they are evidence against themselves varies from the certain self-defeat of 'There are no sentences' to the mildly self-defeating twelve-word sentence, 'At most one sentence on this page is exactly twelve words long'. A second large group of speech acts aim at compliance—a world-to-word fit instead of a word-to-world fit. So we can say an utterance is *normatively* self-defeating when it constitutes an infraction of the imperative it expresses. Unlike 'No sentence is six words long', which counterexamples the proposition it expresses, 'Never write a six-word sentence' disobeys the imperative it expresses. Graffiti writers who scrawl 'Do not write on the wall' on the wall wittingly violate the prohibition but do not falsify it. Archibald MacLeish unwittingly violates his proscription against didactic poetry when he ends his poem "Ars Poetica" with:

A poem should be equal to:
Not true
For all the history of grief
An empty doorway and a maple leaf
For love
The leaning grasses and two lights about the sea—
A poem should not mean
But be

Notice that MacLeish's utterance is normatively self-defeating only as a poetic verse, not as a simple assertion or as the thesis of an argument. The relativity of self-defeat is also illustrated by Elmore Leonard's *52 Pick-up*. In the novel, a blackmail victim learns that his gun killed his mistress by viewing a video tape that was filmed in a way that eliminated the possibility of a fake execution. This is not a self-defeating depiction but the movie version of *52 Pick-up* is; for one scene shows the video tape, which at first does appear unfakeable except when you reflect on the fact that Hollywood does not make snuff films.

In most cases of self-defeat, the infraction is by the speaker. There is, however, a genre of self-defeat in which the audience must break the rule in order to understand it, as with 'Do not read this sentence' and 'Do not think of pink elephants'. Audience participation is also crucial to self-defeating prophecies in which dissemination of the prediction causes a reaction that falsifies it. For instance, the agricultural expert's prediction of corn shortages may prompt farmers to plant more corn.

A "self-defeating argument" can be defined as one that contains a self-defeating premise or conclusion. This allows some arguments to be self-defeating in ways that are independent of their status as argument. For example, an argument for the conclusion, 'No sentence is six words long', contains a self-defeating *assertion*. Incorporating 'There are no arguments', however, yields a self-defeating *argument* even though the statement is not self-defeating as a bare *assertion*. This pure sort of argumentative self-defeat can be normative. Robert Nozick's[13] argument against arguing is normatively self-defeating only because it is an argument; as a bare protest, it does not frustrate itself.

---

[13] *Philosophical Explanations* (Cambridge: Harvard, 1981), pp. 4–5.

The analysis of self-support parallels that of self-defeat; merely substitute 'fulfill' for 'frustrate'. This extension is important because exemplification is openly intended self-support. When an applicant for the job of translator says 'I speak English' to a skeptical interviewer, he intends his assertion as a sample. Ditto for the salesman who argues that he ought to be hired because he is a persuasive speaker, the testifier fending off accusations of insanity, and the logician demonstrating that the study of logic reduces your vulnerability to fallacies. Clever refutations often exemplify what they assert. My favorite was aimed at a linguistic philosopher who was delivering a lecture in which he just noted that the double negative construction in some languages has a positive meaning. He went on to note that, in other languages, the double negative intensifies denial. Yet, in no language, said the lecturer, do we find a double positive construction with a negative meaning. From the back of the lecture room, someone interjected in a jeering voice "Yeah, yeah." A more famous example is Winston Churchill's objection to the prohibition against ending sentences with prepositions: "That is something up with which I will not put." The objection is more than a bare protest because it exemplifies the awkwardness compliance would cause.

Exemplification is at play in a number of philosophical arguments. In the cogito, Descartes's "mental utterance" of 'I think, therefore, I exist' exemplifies his thinking, thereby establishing the truth of the premise. The exemplification feature provides a reply to the objection that Descartes could have just as well used 'I walk', 'I eat', or some other less ponderous assertion as his premise.

More obscure examples can be drawn from a project undertaken by Girolamo Saccheri[14] in 1697. Saccheri aimed to prove the invalidity of invalid syllogisms using rules designed to prove the validity of valid syllogisms. His method can be illustrated with his argument against AEE-1 syllogisms, also called "Maltese" syllogisms.

All Barbara syllogisms are valid.
No Maltese syllogisms are Barbara syllogisms.
———
No Maltese syllogisms are valid.

---

[14] Saccherian syllogisms are discussed in William and Martha Kneale, *The Development of Logic* (New York: Oxford, 1962), pp. 345–8.

The premises are true. So, if the argument is valid, the conclusion must be true. But since the argument is a Maltese syllogism, it would then follow that the argument is invalid. Therefore, Maltese syllogisms are invalid—just as the conclusion asserts. The fascinating feature of Saccheri's argument is that one must recognize its (formal) invalidity in order to be persuaded by it. So it should confound the validity chauvinist.

Many logicians would be leery of Saccheri's syllogism's apparent self-referentiality.[15] But the conclusion does not imply that the argument is itself invalid. That verdict only follows with the ancilliary premise that the argument instantiates the Maltese form. Although this is never *asserted* in the argument, the argument does *exemplify* 'Maltese syllogism'. So the notion of exemplification helps to extricate Saccheri's syllogism from the objection that it embodies a confusion between language and metalanguage.

C. Arguments having logical theorems as their conclusions. My third group of examples is inspired by the premiseless arguments discussed in logic textbooks. The typical conclusion can only be proved by applying inference rules to premises. But if the conclusion is a theorem of logic, the inference rules alone are enough to do the job. Adding a premise to a valid premiseless argument yields a valid single-premise argument whose premise is superfluous. Since we are free to choose any premise, let us take the conclusion itself as the premise, as in:

(I) Either ants spit or it is not the case that ants spit.
Either ants spit or it is not the case that ants spit.

Given that the conclusion is inferred purely by means of inference rules, there is no self-inference, and so no circularity. Yet the argument has the form '*P*, therefore, *P*'.

For (I) to work, it must be understood as an argument with a single, redundant premise. Relative to a solely rule-driven derivation, (I) is not circular. Yet (I) is circular relative to a derivation that makes essential appeal to the premise of (I). As is common practice in logic textbooks, we should distinguish between arguments and proofs. Arguments are not identical to proofs because one argument usually has many alternative proofs of its conclusion.

---

[15] C. L. Hamblin defends Saccherian arguments in "Saccherian Arguments and the Self-Application of Logic," *Australasian Journal of Philosophy*, LIII, 2 (August 1975): 157–64.

D. Examples drawn from pragmatics. Progress in pragmatics (indexicals, implicature, emotive meaning) has overturned the belief that utterances that express the same proposition must have the same meaning and utterances that have the same meaning must express the same propositions. For example, we distinguish between the occasion meaning of 'I bowl' and its constant meaning. Its occasion meaning is the proposition expressed on a particular occasion when we "fill in" the referent of 'I'. When Carter utters 'I bowl', its occasion meaning is that Carter bowls. When Bush utters 'I bowl', its occasion meaning is that Bush bowls. But it also has a meaning that does not vary from speaker to speaker. This constant meaning is not a proposition. Hence, there is a sense in which two statements that do not express the same proposition can mean the same thing. This suggests the possibility of circularity without propositional repetition. For example, pragmatic theories of sarcasm deny that it alters the truth conditions of what is said. The difference it makes to meaning consists of what the speaker suggests. This lends support to the intuitive circularity of

(J) There is a slim chance that the Mets will win.
———————————————————————
There is a fat chance that the Mets will win.

Given that the premise is literal while the conclusion is sarcastic, they express incompatible propositions. Yet they mean the same thing in the sense of conveying doubt that the Mets will win.

In the case of constant meaning, we say that two assertions mean the same thing even though different propositions are expressed. Noncircular '*P*, therefore, *P*' arguments can be generated from the converse case, where we say that two assertions have different meanings even though they express the same proposition. A necessary condition for circularity is that the conclusion repeat something meant in the premise set. So, if the conclusion means something not contained in the premises, the argument is not circular. Once we allow assertions to mean more than the propositions they express, we allow for the possibility that a '*P*, therefore, *P*' argument will have a conclusion that means something different from its premise; for remember that *P* stands for a *proposition* and propositions do not exhaust meaning.

Grice's theory of implicature provides widely accepted evidence that meaning goes beyond the truth conditions of the statement. For instance, the premise and conclusion of the following express the same proposition but have different conversational implicatures:

(K1) Princess Di got married and pregnant.
Princess Di got pregnant and married.

Whereas the premise suggests that Di first married and then later conceived, the conclusion suggests the reverse temporal order. This will lead many of Princess Di's champions to affirm the premise but deny the conclusion. Even defenders who grant that the premise and conclusion express the same proposition will object that the argument is misleading—not that it is circular. Of course, those who hold that '*P* & *Q*' is a different proposition than '*Q* & *P*' will deny that (K1) has the '*P*, therefore, *P*' form. But such fastidiousness about propositional identity would be too picky for anyone who believed that circular reasoning is common in ordinary argumentation; for naturally occurring arguments usually have slight differences between premises and conclusions that "really say the same thing."

The conclusions of the next pair share their premise's truth conditions but differ in pragmatic meaning by an extraconventional implicature.

(K2) Mary is poor and honest.
Mary is poor but honest.

(K3) Tim can answer it.
Even Tim can answer it.

The conclusion of (K2) suggests that poverty tends to preclude honesty. The conclusion of (K3) gives the impression that the question is easy. Once again, I deny that these arguments are circular on the grounds that the premise and conclusions do not mean the same thing.

The divergence between semantic and pragmatic synonymy can be subtle. Highly status-conscious languages such as Japanese extend honorific usage beyond people to their groups, actions, and belongings. English has a pale shadow of this phenomenon in the form of elevation: (residence, home), (dine, eat), (lady, woman), (conjecture, guess), (postulate, assume), (perspire, sweat), (imbibe, drink). The left-hand member of each pair confers higher status than the right without any descriptive supplement. This difference between semantic and pragmatic synonymy is the basis for

(L1) Vernon owns a violin.
Vernon owns a fiddle.

Emotivists were attracted to more strident phenomena: honorific and pejorative terms. According to the emotivists, some expressions that are descriptively synonymous vary in emotive meaning.

(L2) Mary is pig-headed.
Mary is resolute.

(L3) Joe is a German.
Joe is a Kraut.

In (L2), we move from pejorative to honorific. In (L3), the premise lacks emotive meaning but is descriptively equivalent to its charged conclusion.[16]

Given pragmatic theories of emphasis and focus, these also work:

(M1) Hester is the winner.
Hester is *the* winner.

(M2) Good food is not cheap food.
Cheap food is not good food.

Argument (M1) brings the uniqueness condition of definite descriptions to the fore by emphasis. In (M2), the premise brings to mind good, expensive food, while the conclusion prompts thoughts of cheap, bad food. Once again, disagreement about propositional identity may lead some to deny (M2) has the '*P*, therefore, *P*' form on the grounds that converses are different propositions rather than different expressions of the same proposition.

The last examples are inspired by pragmatic theories of metaphor.[17] These theories say that metaphorical statements do not differ in truth conditions from their literal counterparts. Read the premise of the (M3) literally and the conclusion metaphorically, and do the reverse for (M4).

(M3) Tons of fish were in the boat.
Tons of fish were in the boat.

(M4) Mayor Koch has a big mouth.
Mayor Koch has a big mouth.

One can imagine the premise of (M3) used to justify the metaphorical conclusion. A person with a penchant for "resemblance thinking" might use the metaphorical premise of (M4) to justify his anatomical conclusion.

E. Aspect cases. Perhaps propositions are like gems. Just as the same gem can have many interesting facets, a proposition might have different aspects that vary in salience. Bearing this analogy in mind, read the following as an inference from a passive to an active change:

---

[16] J. O. Urmson has a good discussion of emotive meaning in *The Emotive Theory of Ethics* (London: Hutchenson, 1968), ch. 10.

[17] Donald Davidson defends a pragmatic theory of metaphor in "What Metaphors Mean," in *On Metaphor*, Sheldon Sacks, ed. (Chicago: University Press, 1979), pp. 29–45. James D. Carney also defends it in "The Meaning of Metaphor," *Philosophical Studies*, XLIV, 2 (September 1983): 257–67.

(N1) The newspaper acquired new owners.
New owners acquired the newspaper.

Although the premise and conclusion have the same logical form, the premise highlights the newspaper's continuity. The slight orthographic difference between the premise and the conclusion might lead some to suppose that different propositions are involved. This defect is eliminated by the following example, which should be read as an inference from the existence of tenants to the existence of landlords:

(N2) Someone rents a house.
Someone rents a house.

If distinct propositions are involved, then the example is as bogus as the ambiguous ones discussed in the second section. Yet note that the premise and conclusion are orthographically identical and strictly entail each other. The felt difference between the readings could be explained as one of aspect. People attend to the existence of tenants when they first hear 'Someone rents a house' because tenants are more common than landlords. But the proposition also entails the existence of a landlord. The difference in salience makes belief in the landlord lag behind belief in the tenant. But redundancy prompts a second look, allowing the laggard to catch up.

IV. ARE 'CIRCULAR' AND 'QUESTION-BEGGING' SYNONYMOUS?

There are at least two commentators who would object to the fact that I use 'circularity' and 'question-begging' interchangeably. Robert Fogelin's[18] logic text contains the following passage:

> Closely related to the fallacy of circular reasoning is the subtle fallacy called *begging the question*. In circular reasoning, the conclusion itself or a restatement of the conclusion in different words is used as a premise in the argument. In begging the question, a statement is made that presupposes or depends upon the point at issue (*ibid.*, p. 95).

Other writers, such as Kent Wilson, reserve 'question-begging' for arguments that are *viciously* circular. Although the synonymy deniers differ on how the meanings diverge, they will be tempted to cast my dispute with the syntactic theorist as merely verbal; for one might craft the distinction to reach the diplomatic solution that my '*P*, therefore, *P*' arguments are circular but not question-begging. 'Circular' might be defined as a purely syntactic concept and 'question-begging' as a pragmatic or semantic concept catching the residual nonsyntactic sorts of dependencies.

---

[18] *Understanding Arguments* (New York: Harcourt Brace & Jovanovich, 1987).

Etymologists tell us that synonyms are unstable; one synonym is either dropped or changes its meaning to cover a nuance. My opinion is that Fogelin and Wilson are pressuring 'circular' and 'question-begging' into the latter fate. In any case, the synonymy issue is a red herring; for even if the debaters grant that there is a difference in meaning, they will only reformulate the debate in the new terminology. The syntactic theorist will deny that he is only analyzing the diplomat's concept of circularity. His goal was to achieve a formal understanding or explanation of an *interesting* fallacy. Confining the meaning of 'circular' in a way that obviously guarantees the adequacy of a syntactic account attains tranquility at the price of triviality. So, if the syntactic theorist accepts the distinction, he will reformulate his thesis so that his enterprise encompasses question-begging. On the other hand, if 'circular' is not defined in a way that makes him obviously right, we lose confidence that the dispute is verbal. So my reply to the objection that 'circular' and 'question-begging' are not synonymous is (a) *reportive* definitions will show their synonymy and (b), even if they diverge, the difference will be irrelevant to my thesis.

V. CONCLUDING REMARKS

The noncircular '*P*, therefore, *P*' arguments are intended to be mutually supportive even though they are logically independent. Once you accept one example, resistance is broken and you are apt to accept many more examples.

The examples are more than curios. First, they are relevant to a controversy over how circularity is to be analyzed. Noncircular '*P*, therefore, *P*' arguments refute the traditional syntactic account of circularity and confirm the pragmatic approach in a way unanticipated by the pragmatists. Second, some of these examples suggest that there is a neglected parameter of evaluation: exemplification. Those who accept the suggestion will greet the cases as more than logic teasers.

ROY A. SORENSEN
New York University

BOOK REVIEWS

History, Labour and Freedom: Themes from Marx. GERALD A. COHEN. New York: Oxford University Press, 1988. xiii, 317 p. Cloth $65.00, paper $24.95.*

History, Labour and Freedom (HLF) is comprised of fourteen essays by G. A. Cohen, based on papers published, with two exceptions, after the appearance of *Karl Marx's Theory of History: A Defence* (KMTH).[1] Six essays have been extensively revised or reorganized, four are reprinted virtually unchanged, and four have been modified slightly. Five of the papers in this collection continue the project begun in KMTH: the exposition and defense of historical materialism. There are also four essays in which Cohen himself takes aim at the theory defended in KMTH. Of the five remaining pieces, three deal with freedom; one with exploitation and the labor theory of value; and one, "The Dialectic of Labour in Marx," originally published fifteen years ago, grapples with notions of alienation and dialectics, themes largely ignored in analytical Marxian writings of more recent vintage. Cohen's papers on self-ownership and equality are not represented in this collection; nor is there much indication of his long-standing engagement with libertarian social philosophy.

Because it collects and updates Cohen's recent work on historical materialism, HLF is a convenient and timely addition to the literature KMTH has spawned. Cohen's position in this on-going discussion has become increasingly ambivalent. By his own account, he continues to defend the cogency of the theory developed in KMTH, at the same time that he has become skeptical of its truth. An uncharitable reader might conclude that he defends his reconstruction of the theory against all critics, while advancing hesitations of his own about its plausibility and importance. But however Cohen's stance is construed, the cogency of the arguments he adduces makes his recent positions, pro and con, as nearly indispensable to this debate as KMTH itself.

HLF is also useful for drawing together some of Cohen's contributions to the analysis of freedom, his critique of the labor theory of value, and, in the essay on the dialectic of labor, his account of Marx's intellectual debt to Hegelian philosophy. But for readers

---

* I am grateful to G. A. Cohen, Daniel Hausman, Debra Satz, and Erik Olin Wright for comments on an earlier draft of this review.

[1] Princeton: University Press, 1978. Published simultaneously in Great Britain by Oxford University Press.

0022-362X/91/8805/267–75

interested in Cohen's work on normative issues arising out of his penchant for taking libertarianism seriously, this collection will only whet the appetite. There is need for a second volume of essays responsive to this interest.

As in KMTH, Cohen's reconstructions and critical assessments of Marx's views consistently conform to the highest standards of argumentation and scholarship. Many positions of varying degrees of contentiousness are staked out and defended with great acuity. These essays therefore invite engagement informed by the conviction that, in philosophy and scholarship, "God is in the details." Inevitably, there are, in the details, ample opportunities to quarrel with Cohen's views. But there are also general themes that emerge, particularly in the essays on historical materialism, that call for consideration in their own right. At this point in the development of the debate KMTH originated, it is instructive to address Cohen's recent work at this level. I shall therefore focus on one of the themes advanced in this volume: the claim that historical materialism, if sustainable, explains less than Marxists have traditionally supposed, and the related idea that the theory matters very little for other Marxian theoretical endeavors or for Marxian politics. I shall take issue with Cohen on both counts: suggesting, first, that the explanatory pretensions of Marx's theory be diminished even more radically than Cohen proposes, but then insisting that, even so reduced, historical materialism plays a more important role in the Marxian project than Cohen maintains.

I

According to the theory reconstructed and defended in KMTH, human history—from the beginnings of class society until the advent of communism—has a determinate structure and direction, in consequence of an endogenous dynamic process in which social "relations of production" rise and fall in order to maintain functional compatibility with developing material "forces of production." Marx sketched this position in a few brief paragraphs in the 1859 Preface to *The Critique of Political Economy*, a text that has become canonical for "orthodox" historical materialists. No one since Marx has done more than Cohen—in KMTH and in some of the essays in HLF—to elaborate and develop the general schema announced in 1859: explaining what forces and relations of production are, how Marx conceived their interaction, and why forces tend to develop continuously, while sets of production relations change discontinuously to accommodate ever increasing levels of development of productive forces.

It might appear that the internal dynamic Marx postulated would seldom, if ever, figure in explanations proffered by practicing historians, because a historian's *explanandum* is typically more fine-grained than a historical materialist's. Marx's theory of history does not, after all, purport to explain *events*. It is a theory of history itself, of its (epochal) structures and of the processes that impel movement from one structure to another. Mainstream historiography assumes that history in this sense has no theory: that historical events, no matter how individuated, may all be explicable at least in principle, but that even in principle there exists no explanation for history itself—except perhaps in the trivial sense achieved by conjoining all explanations of particular events. For the explanatory objectives of practicing historians, it would therefore seem that Marx's theory of history would be of little interest. Arguably, the apparent irrelevance of historical materialism for Marx's own historical writings attests to the soundness of this inference.

When Cohen suggests that the Marxian project can survive a weakening of historical materialism's explanatory pretensions or even outright abandonment of historical materialism itself, his rationale does not directly appeal to historical materialism's ostensibly superfluous role in historical explanations, but to its irrelevance for normative criticisms of capitalist society. He writes:

> Some people may think that, in raising a challenge to historical materialism, I am letting the side down. So let me emphasize that my belated reservations about the theory do not weaken my belief that it is both desirable and possible to extinguish existing capitalist social relations and to reorganize society on a just and humane basis. The political significance of retreat from historical materialism should not be exaggerated. An appreciation of the principal evils of capitalism, which are its injustice, its hostility to the development of the faculties of the individual, and its voracious ravaging of the natural and built environment, does not depend on ambitious theses about the whole of human history. Nor does the claim that it is possible to establish a society without exploitation which is hospitable to human fulfillment require, or perhaps even follow from, those theses. So scepticism about historical materialism should leave the socialist project more or less where it would otherwise be (132).

I would suggest, however, that more is at stake in debates about Marx's theory of history than Cohen suggests in this passage; and that failure to appreciate historical materialism's foundational role in Marxian theory and politics can lead to an unnecessarily hasty rejection of Marxism itself.

The historical materialist orthodoxy Cohen reconstructed and defended in KMTH posits functional relations between forces and relations of production, and between sets of production relations or "economic structures" and legal and political "superstructures" and forms of consciousness. The most trenchant of the reproaches Cohen now levels against the theory of KMTH or, more precisely, against the traditional understanding of its explanatory scope, pertains to the second of these functional relations. On the traditional "inclusivist" view, economic structures explain everything—at some suitable (but unspecified) level of generality. What Cohen now suggests is that historical materialists need only hold that economic structures explain superstructural facts to the extent that is necessary to maintain the autonomy of the endogenous dynamic process that causes economic structures to rise and fall. For a "restricted historical materialist," noneconomic facts cannot impinge on this underlying process in the wrong way; they cannot upset its dynamic. But they can fail to affect that process at all. Insofar as they remain causally aloof, historical materialists have no reason to maintain that they are in any way explained by the process the theory identifies. Orthodox historical materialism was a theory of general history—that is, of virtually everything historical. In suggesting that Marxists have reason to be restricted, not inclusive, historical materialists, Cohen, while enhancing its plausibility, radically deflates historical materialism's explanatory pretensions. Restricted historical materialism would hardly count as a theory of general history. But it is a foundational theory nevertheless.

There is a second reproach Cohen levels against historical materialist orthodoxy that addresses the underlying dynamic it purports to identify. That process is propelled by a transhistorical human interest in developing productive forces. In several of the essays in this volume, Cohen defends this claim against charges of circularity. But then he advances other reasons for disputing its putative role in moving history along. He suggests that the historical dynamic may be shaped as well by other interests—particularly, as Hegel believed, by an interest in "self-identification." If so, historical materialism may fail not just as a theory of general history, but also as a general theory of history—that is, of historical trends—precisely because it is fatally one-sided.

Defenders of historical materialism can resist this conclusion, however, and still concede that an interest in developing productive forces is not the only properly transhistorical interest shaping human mentalities. Historical materialists can even concede the existence of a dialectic of self-identification of the Hegelian sort and invest it with historical efficacy, as Cohen suggests. To see why, it will help to follow Cohen's own lead by focusing just on what the theory can plausibly be thought to explain, regardless of the explanatory ambitions its adherents have traditionally assumed.

The historical materialism of KMTH purports to provide an account of necessary and sufficient conditions for epochal historical transformations—from one set of production relations to another. It holds that certain levels of development of productive forces are necessary for the inception and reproduction of particular "economic structures" or sets of production relations and that continuing development generates "contradictions" (structural instabilities) between forces and already existing relations of production. Thus, in the orthodox view, the development of productive forces is sufficient in the long run for ensuring movement from one set of production relations to another. It is only when productive forces are massively developed and harmony between forces and relations of production is ensured under socialism that this process finally ceases to govern the transformation of economic structures.

It is not immediately apparent what would count as evidence for or against the existence of this internal dynamic. But on a superficial reading of human history, it would seem to explain little, if anything, of what has actually transpired. The historical-materialist dynamic should figure prominently, for example, in explanations for the emergence of capitalism. But in England and the parts of Western Europe where capitalism first became the dominant economic structure—the one instance in which it was not introduced by (imitative) domestic elites or imposed by outside powers—even some Marxian historians accord explanatory priority to such factors as peculiarities of class structure, the pattern of town-countryside relations, the nature of political institutions, and demographic considerations. The fettering of productive forces under feudal relations of production does not seem to play an important explanatory role.[2] It would therefore seem that the orthodox account of necessary and sufficient conditions for change, like the orthodox view of the relation between base and superstructure, is embarrassingly susceptible to disconfirmation. The idea that the theory is misleadingly one-sided would explain this unhappy situation. But in the apparent absence of an analogue to restricted historical materialism to which to retreat, the implication is not that the theory be weakened in its explanatory ambitions, but that it be overthrown altogether. This appears to be the conclusion toward which Cohen now tends. If Marx was wrong to invest the dynamic of history exclusively in an interest in developing productive forces, the theory that follows from this premise would be flawed at its core. No wonder, then, that it should square poorly with the facts of the matter. But, *pace* Cohen, it is not necessary to destroy the theory in order to save it. There are less radical ways to shelter Marx's account of history's underlying dynamic—or, more precisely, the core insight that motivates it—from the likelihood of empirical confutation.

---

[2] Not surprisingly, debates about the origins of capitalism have long preoccupied Marxists. Several decades ago, this debate was waged in expressly Marxian terms in the pages of *Science and Society* in a series of articles sparked by Paul Sweezy's criticisms of Maurice Dobb's *Studies in the Development of Capitalism* (New York: International, 1947, repr. 1963, 1972). The *Science and Society* debate has been republished, with additional material, as *The Transition from Feudalism to Capitalism* (New York: International, 1976). Since the publication of Robert Brenner's article, "Agrarian Class Structure and Economic Development in Pre-Industrial Europe," *Past and Present*, LXX (February 1976): 30–75, the Marxian debate over the transition from feudalism to capitalism has been

One could, for instance, rely on the distinction between the general laws historical materialism advances and the hypothesis that these laws actually do play a prominent role in explaining real history.[3] If the case for historical materialism developed in KMTH is cogent, then there is good reason to believe in the reality of an endogenous process capable of moving history from one epochal structure to another. In the absence of other pertinent causal factors, this process would explain what actually transpires. But, of course, other factors are always present. Thus, those who account for, say, the emergence of capitalism by appeal to considerations distinct from the ones historical materialism postulates need only maintain that the causes they invoke swamp the historical-materialist dynamic; not that the process historical materialists identify does not exist. Then the theory of KMTH would be explanatory, but much less so than Marxists have traditionally supposed.

There is, however, a more plausible way to regard the Marxian theory of history which recalls the retreat Cohen himself proposed

---

joined by historians of a variety of theoretical orientations—in the form of competing explanations for different levels of economic performance in different societies in early modern Europe. Some of the principal contributions to this debate are collected in T. H. Aston and C. H. E. Philpin, eds., *The Brenner Debate: Agrarian Class Structure and Economic Development in Pre-Industrial Europe* (New York: Cambridge, 1985).

[3] Cf. Andrew Levine and Elliott Sober, "What's Historical about Historical Materialism?" this JOURNAL, LXXXII, 6 (June 1985): 304–26, esp. pp. 312–4.

in distinguishing restricted from inclusive historical materialism. This proposal ill accords with Marx's own understanding of the theory he developed or with the orthodoxy reconstructed in KMTH. But it does advance Marx's "rational intuition" in a plausible way—at the same time that it reveals, as perspicuously as even the most stringent orthodoxy, historical materialism's foundational role in Marxian theory. I have elsewhere called the sort of theory I have in mind *weak historical materialism.*[4] Weak historical materialism does not advance sufficient conditions for epochal historical change, but only necessary material conditions. It does not purport to explain what actually happens, other things being equal (or nonexistent), but only what is (materially) possible. By identifying correspondences and "contradictions" between forces and relations of production, it specifies a map of historically possible epochal structures, and also the possibilities for moving from one epochal structure to another. In short, it establishes a determinate *historical agenda.*

II

Whether in its weak or traditional form, historical materialism identifies social classes as the agents who move history along. In the Marxian scheme, social relations of production constitute class positions into which individuals fit; and it is the concerted activities of individuals, as class actors, who reproduce—or overthrow and establish—the economic structures that comprise history's epochal divisions. Thus, historical materialism forces attention on class structure, class relations, and class struggle. To be sure, proponents of weak historical materialism—particularly if they are also restricted historical materialists—are bound to endorse class analysis only insofar as the phenomena they seek to explain bear on the historical-materialist dynamic. But this commitment is enough to motivate a distinctive research agenda.

Understood in the usual way, as a claim about what explains particular phenomena, it is almost certainly false that class is always more explanatory than other factors. But the usual understanding is misleading. What drives the Marxian worldview is the idea that power—at the level of social structure—accounts for social order and change; and that real (as opposed to merely juridical) ownership of productive resources, the basis for class divisions in the Marxian sense, is the source of this power. Explicit class identifications or overt class conflicts need not have more explanatory force in particular instances than other considerations. But at the level of generality appropriate for understanding epochal change, class relations render social life intelligible. For Marx, intelligibility at this level is indispensable for a political practice directed toward the overthrow of capitalism and its replacement by socialism, its historical alternative and successor.

---

[4] See *ibid.*, pp. 323–6, and my *The End of the State* (London: Verso, 1987), ch. 5.

Thus, despite Cohen's imperturbability, Marxists cannot so easily do without Marx's theory of history. The normative critique of capitalism Marxists proffer would indeed survive historical materialism's demise. But if moral criticism were the only theoretical resource propelling Marxists' struggles for socialism, Marxism would devolve into a utopian socialism of the kind against which Marx vehemently inveighed. Marxists have insisted repeatedly that their socialism is "scientific," not utopian, precisely because it is grounded in an account of real, material processes—not moral ideals, however compelling. Historical materialism, even in its restricted and weak form, is crucial for sustaining this claim. Marxism is not a utopian socialism precisely because its theoretical and political project is grounded in a theory of history's structure and direction.[5]

No doubt, Marx and his successors were carried away by their polemics against utopian socialism. They had no need to oppose, as they sometimes did, the very idea of normative theory. Indeed, Cohen has shown—no where more persuasively than in an essay included in this volume, "Historical Inevitability and Revolutionary Agency"—that orthodox historical materialism is compatible with and even requires moral criticism to motivate revolutionary agents. If historical materialism, properly revised, is not a theory of historical inevitability after all, the need for a clear perception of the evils of capitalism becomes all the more pressing. It is therefore urgent that socialists take back on board the moral imperative driving the utopian-socialist vision. Many of the essays in this collection contribute importantly to this task. But, in the final analysis, for just the reasons Marx adduced, it is not enough that socialism be morally necessitated. Marx insisted that socialism be inscribed in a historical dynamic. The historical materialism of KMTH, if sustainable, would ensure that it is. But it is not clear why "scientific" (nonutopian) socialist theory must endorse socialism's inevitability. It should be enough to defend its place on the historical agenda. This is precisely what weak historical materialism does.

---

[5] There, of course, remains the possibility that some rival account of history's structure and direction could do the work historical materialism does in the Marxian scheme. Perhaps, as Cohen suggests, a "dialectic of self-identification" might play some role in such a theory. At the present time, however, this proposal is only programmatic. Historical materialism is surely not the only imaginable theory capable of grounding the theoretical and political project Marx envisioned, but it is, for now, the only full-fledged theory extant.

The question of the viability of historical materialism—or of successor theories that retain its core insights—should therefore be an issue of great urgency for anyone concerned with Marxism. In moving from inclusive to restricted and then from traditional to weak historical materialism, the plausibility of the theory is enhanced, without detriment to its foundational role in class analysis and class politics. But is what remains true? No considerations so far adduced by Cohen or anyone else require us to conclude that it is not. But neither, at this point, can we be sure that the core insight historical materialism articulates can finally be sustained.

Whatever his current hesitations, Cohen remains the premier defender of Marx's theory. This is why his suggestion that historical materialism may ultimately prove false cannot fail to disturb. I think Cohen would agree that the jury is still out. He should also agree that the issue cannot be settled by philosophical argument alone. Since the aim of the theory is to make sense of history at a certain level of generality, historical evidence and judgment must ultimately adjudicate the theory's claims. But philosophers can help by clarifying what corroboration or disconfirmation involves. Cohen's recent work—even at its most skeptical—cogently addresses this complex and still scarcely examined issue. Not least among the merits of the essays on historical materialism included in this volume, therefore, is the fact that they provide a point of departure for this next stage in the investigation of Marx's theory of history: an inquiry into the empirical conditions for its tenability.

ANDREW LEVINE

University of Wisconsin/Madison

The Shaky Game: Einstein, Realism, and the Quantum Theory. ARTHUR FINE. Chicago: University of Chicago Press, 1986. 186 p. Cloth $25.00.

What exactly were Einstein's attitudes toward quantum theory? Do they add up to no more than a naive realism which we can no longer

0022-362X/91/8805/275–9

share? More generally, is scientific realism a defensible philosophical position? These are the questions which Arthur Fine addresses in *The Shaky Game*. The book consists of a short introductory chapter followed by eight essays written over a ten-year period. Essays 2-6 are concerned with Einstein's views on quantum theory, as they developed during the 1920s and 1930s in opposition to prevalent Copenhagen attitudes. Then, in the last three essays of the book, Fine discusses the general issue of scientific realism, and advocates a stance—he calls it the *natural ontological attitude*, or NOA—free of the rigidities he associates with realism, on the one hand, and various versions of antirealism, on the other.

As early as 1926, Fine tells us, Einstein voiced his concern about the following features of Schrödinger's theory. (1) Its equations are not relativistically invariant. (2) It does not yield classical behavior to a good approximation. (3) It leads to correlations between distant objects that violate action-by-contact principles. (4) It is an essentially statistical theory. (5) The scope of the commutation relations may not be as broad as the theory suggests. In particular we may note item (3), and that item (5) was aired *before* Heisenberg had formulated his uncertainty relations. Fine suggests that, after the Solvay conferences in the late 1920s and his discussions with Bohr, Einstein accepted that the commutation relations were indeed built into the theory; that reduced the list to four issues, concerning which Einstein was never satisfied.

In 1935, jointly with Podolsky and Rosen, Einstein published the paper, "Can Quantum Mechanical Description of Reality be Considered Complete?" (now usually known as EPR). In the same year there appeared the pair of papers by Schrödinger that introduced us to that legendary beast, Schrödinger's cat, and thereby showed how counterintuitive was the application of quantum principles to the macroscopic world. These papers are still required reading for anyone studying the foundations of quantum theory. In one of the most interesting chapters in *The Shaky Game*, Fine draws on the correspondence between Einstein and Schrödinger to reconstruct the way in which the Schrödinger-cat papers took shape, and to relate them to the EPR paper. This essay is a splendid piece of historical reconstruction.

Again, by looking at sources other than the published version of EPR (which was written by Podolsky), Fine shows that, for Einstein, the thought experiment it described had less to do with the inability of quantum mechanics to specify all "elements of reality" than with the dilemma that the theory poses: *either* a principle of separation fails to hold, *or* the theory is not complete. Thus the "principle of separation"—that the "real state" of one system cannot depend on what sort of measurement is carried out on a spatially separated system—was seen by Einstein as a key issue long before Bell made nonlocality a hot topic.

Einstein advocated a *statistical interpretation* of quantum theory, according to which the wave functions appearing in the theory describe, not single systems, but ensembles of systems. On this topic Fine makes two suggestions. Both are interesting, and both find support within Einstein's writings. There is, however, a tension between them, and my one reservation about the first part of Fine's book is that he resolves it in too facile a manner. The first suggestion is that Einstein was not interested in a hidden-variable theory, as that term is generally understood. That is, he was not interested in a supplementation of quantum mechanics *from within*. Rather, he looked for a recasting of theory that would attend to the "external" problems that quantum mechanics faces; he sought a theory which would mesh with relativity, which would have no awkward quantum/classical dichotomy built into it, and so on. One desideratum of such a theory was that in it the quantum mechanical statistics would appear as averages over ensembles, but that did not mean that the observables with which quantum theory deals were to have a special status.

Fine's other suggestion is that Einstein could escape the problems that Bell's theorem poses for ensemble interpretations by the use of *prism models*. Prism models were designed by Fine to deal with the fact that quantum mechanical probabilities cannot readily be inscribed within a classical probability space. On this approach, the probabilities associated with each measurement procedure are seen as averages over a particular subensemble of the systems tested, rather than over the whole ensemble; the whole set of quantum mechanical probabilities is then obtained by piecing together classical probability measures over different, but overlapping, subensembles.

My suspicion, however, is that the move to prism models would leave Einstein cold; it has the ring of a formal, rather than a physically motivated move. Fine acknowledges as much. Although the prism models he describes jibe with various things Einstein says—and says repeatedly—about ensemble interpretations, what they offer is effectively a hidden-variable theory, albeit one immune from Bell's challenge. As Fine says, they are clearly supplementations of the theory from within.

On the one hand, then, Einstein demands a reconstitution of physical theory, which to date remains set as a task; on the other, he speaks in ways that seem to invite the kind of supplementation that prism models provide. I see this as a genuine tension in Einstein's thought, not one that is introduced by Fine's exegesis. Indeed, Fine does a remarkable job of sifting Einstein's rather sparse and laconic utterances on these matters to reveal the two tendencies within them. Where he overstates the case on Einstein's behalf, I think, is in his resolution of the tension. When Einstein talked of an "ensemble interpretation," Fine suggests, he "chose his rhetoric cunningly" (58). In other words, those passages where he seems to be advocating, if not a prism-model approach, then one very like it, are not to be taken at face value. On the documentary evidence Fine himself presents, this seems implausible. Perhaps in this instance we should allow Einstein the luxury of being in two minds. Maybe there was an ensemble interpretation that would solve the local problems within quantum theory; ultimately, however, quantum theory would have to answer to more global demands.

In the last three essays of the book, Fine addresses the question of scientific realism. The opening sentence of essay 7, 'Realism is dead', and the title of essay 8, "And Not Anti-Realism Either," announce the dialectic of the debate. Essay 9 considers whether quantum mechanics, or any physical theory, rules out scientific realism. Fine argues, and I agree with him, that it cannot, but that the realism it permits is of a minimal kind. In what follows, I shall focus on essays 7 and 8.

The realism whose obituary appears in essay 7 is the view that the entities postulated by a successful science really exist, and that, when a branch of science consistently provides explanations of the kind we value, we should regard it as true, or approximately so; truth here is thought of as correspondence with reality. For support, realism looks to the success of scientific practice. This can only be explained, the argument runs, on the assumption that realism is true. But, as Fine points out, this argument relies on the very point at issue, to wit, whether the explanatory efficacy of a hypothesis is a warrant of its truth; for this reason, it will hardly convince the skeptic.

Fine does not, however, embrace antirealism. The varieties of antirealism he describes replace the realist's correspondence theory of truth with others: pragmatic, reductive, and so on. In contrast, Fine rejects the search for a theory of truth; we know the history, logic, and grammar of truth, he tells us, and do not need to "explain" it or to "give an account of it." Instead, the attitude he recommends (indeed, the NOA) is to say that, yes, there are electrons, but not to offer a "deeper" account of what that statement says. The NOA-er (or *knower*) simply agrees that electrons exist, but neither adds 'really' to emphasize a correspondence with some fundamental reality, nor glosses the phrase by saying, 'But, in using electron-talk, I simply mean . . .'. To ask the NOA-er what the existence claim amounts to would be to miss the point.

That is one reason why NOA is difficult to pin down. ("If I could understand it," one thinks, "I might accept it." The NOA-er, like a good Zen master, replies, "There is nothing to understand.") Another reason is that NOA has two aspects. Fine wants philosophers to reject the heuristic view that they can provide a unified account, whether realist or antirealist, of the aims, procedures, and achievements of science: the "performance" of science, he writes (148), "is not susceptible to a reading or interpretation in any global sense." We should instead let science speak for itself; our epistemic attitudes toward, for example, magnetic monopoles should be "governed by the very same standards of evidence and inference that are tolerated by science itself" (150).

We are given two prescriptions, one negative, the other positive. But the two by no means come to the same thing. In the first place, it is possible to maintain interpretive flexibility even while criticizing science from without. Fine's own book shows as much. In the second, the voices of realism and antirealism are heard, sometimes stridently, within science itself.

This brings me to my last point. Fine has given us a very good book; the historical essays are marvelously illuminating, and the arguments in the later essays are both subtle and challenging. Yet, probably because these essays reflect the movement of Fine's thought over a number of years, the two parts of the book lie somewhat uneasily together. Just as we can ask whether Socrates could find a home in the Republic, we may wonder about Einstein's fate in a world where the NOA was the one prescribed.

R. I. G. HUGHES

University of South Carolina

NEW BOOKS: TRANSLATIONS

DELEUZE, GILLES: *Expressionism in Philosophy: Spinoza.* MARTIN JOUGHIN, trans. New York: Zone, 1990. 445 p. Cloth $24.50.

HEIDEGGER, MARTIN: *Kant and the Problem of Metaphysics:* Fourth Edition, Enlarged. RICHARD TAFT, trans. Bloomington: Indiana UP, 1990. 224 p. Cloth $35.00.

IRIGARAY, LUCE: *Marine Lover of Friedrich Nietzsche.* GILLIAN C. GILL, trans. New York: Columbia, 1991. 190 p. Cloth $35.00.

KIERKEGAARD, SØREN: *Eighteen Upbuilding Discourses.* HOWARD V. HONG and EDNA H. HONG, eds. and trans. Princeton: University Press, 1990. xxii, 559 p. Cloth $59.50.

KRÄMER, HANS JOACHIM: *Plato and the Foundations of Metaphysics: A Work on the Theory of the Principles and Unwritten Doctrines of Plato with a Collection of the Fundamental Documents.* JOHN R. CATAN, ed. & trans. Albany: SUNY, 1990. xxx, 311 p. Cloth $49.50, paper $16.95.

LEVINAS, EMMANUEL: *Nine Talmudic Readings.* ANNETTE ARONOWICZ, trans. Bloomington: Indiana UP, 1990. xxxix, 197 p. Cloth $29.95.

NEGRI, ANTONIO: *The Savage Anomaly: The Power of Spinoza's Metaphysics and Politics.* MICHAEL HARDT, trans. Minneapolis: Minnesota UP, 1991. xxiii, 277 p. Cloth $44.95, paper $18.95.

RICKEN, FRIEDO: *Philosophy of the Ancients.* ERIC WATKINS, trans. Notre Dame: University Press, 1991. x, 280 p. Cloth $24.95, paper $12.95.

VICO, GIAMBATTISTA: *On the Study Methods of Our Time.* ELIO GIANTURCO, trans. With *The Academies and the Relation Between Philosophy and Eloquence*, DONALD PHILLIP VERENE, trans. Ithaca: Cornell, 1990. xlix, 90 p. Cloth $24.95, paper $6.95.

THE JOURNAL OF PHILOSOPHY
is pleased to announce the publication of its
TWENTY-FIVE YEAR INDEX, 1964–1988
VOLUME LXI TO VOLUME LXXXV
AUTHORS, SUBJECTS, & REVIEWS
Supplement to its
FIFTY-YEAR INDEX, 1904–1953
TEN-YEAR SUPPLEMENT, 1954–1963

It is available for the rest of 1991 at special prices:

25-year (1964–1988) $25.00 cloth/$10.00 paper

also available

50-year (1904–1953) $12.00 cloth 10-year (1954–1963) $3.50 paper

Shipping $5.00 (first)/$2.50 (each addl.)

Please send a check or money order to:

Journal of Philosophy
704 Philosophy Hall
Columbia University
New York, New York 10027

Prices are in effect through December 31, 1991. Orders and payment must be received by then.

THE JOURNAL OF PHILOSOPHY

VOLUME LXXXVIII, NO. 6, JUNE 1991

CONTRACTUALISM, MORAL MOTIVATION, AND PRACTICAL REASON*

John Rawls suggests in *A Theory of Justice*[1] that the contractarian idea might be extended to (1) explicate the concept of right, and (2) argue for "the choice of more or less an entire ethical system . . . including principles for all the virtues and not only for justice" (*ibid.*, pp. 17, 111; cf. pp. 184, 584–5). T. M. Scanlon[2] has outlined an account that in many ways responds to Rawls's suggestions. Contractualism, he contends, offers (1) a "foundational account" of the subject matter of morality, and (2) a general moral view. Scanlon's account is incomplete. But it presents a conception of morality many will find compelling as an alternative to both noncognitivism and utilitarianism. Especially appealing is the contractualist account of moral motivation, since it provides a way to understand obscure but central aspects of Kantian views: the claims that moral reasons are of a special kind, and that moral motives have a basis in practical reason. My aim is to develop Scanlon's account to show how it exhibits these features. Contractualism's primary significance lies in the way it relates motivation to practical reason. This is the subject of sections II–III. To provide necessary background, I begin with a discussion of Scanlon's account. He presents contractualism mainly by contrast with the philosophical bases of utilitarianism. But to appreciate how distinctive contractualism is requires its comparison with other foundational and normative views.

---

* I am indebted to T. M. Scanlon, Burton Dreben, and Jay Wallace for helpful comments on drafts of this paper. My greatest debt is to John Rawls, especially to his unpublished lectures on Kant's ethics. A research grant from the University of Pennsylvania enabled me to write this article.

[1] Cambridge: Harvard, 1971. Hereafter cited as TJ.

[2] "Contractualism and Utilitarianism," in *Utilitarianism and Beyond*, Amartya Sen and Bernard Williams, eds. (New York: Cambridge, 1982), pp. 103–28. Hereafter cited as CU.

0022-362X/91/8806/281–303

I. CONTRACTUALISM AND THE FOUNDATION OF MORALITY

Characteristic of all social-contract views are the ideas that social rules and institutions are to be freely acceptable to all persons bound by them, and for their mutual benefit. These intuitive notions give rise to the idea that moral requirements are grounded in a general agreement. There are different ways to arrive at and describe free acceptability and mutual benefit. In Rawls, they follow from his specification of a conception of social cooperation and a democratic conception of citizens as free and equal moral persons. David Gauthier interprets these ideas starting from an account of rationality as utility maximization, and cooperation-free interaction in a state of nature. As much as these contract views differ, both are directed toward developing a substantive account of political morality. Scanlon's perspective on these ideas (free acceptability, mutual benefit, and general agreement) is different. He sees them, in the first instance, as part of a philosophical account of "the nature of morality" (CU, 110).

Any normative conception of ethics must provide, Scanlon contends, a response to three fundamental philosophical questions:

(1) A foundational question: What is the subject matter of morality, and the grounds of truth, if any, for moral judgments?
(2) An epistemological question: How can we come to know these grounds, i.e., ascertain the kinds of facts that moral truths are about?
(3) A motivational question: Why should anyone care about morality, and be moved to act on its demands (CU, 104–6)?

Utilitarianism derives much of its force, Scanlon says, from the fact that it relies on philosophical premises that give relatively clear-cut answers to these questions (CU, 108). Scanlon calls this view "philosophical utilitarianism," to distinguish it from "normative utilitarian" conceptions, all of which hold the substantive moral thesis that it is right to maximize the sum of well-being. Philosophical utilitarianism is basically a claim about the subject matter of morality; it is "the thesis that the only fundamental moral facts are facts about individual well-being" (CU, 108).

This thesis should not be confused with the idea that morality must advance human interests, their well-being, or their good. These claims are endorsed by a wide range of views—not just utilitarianism and perfectionism, but also Rawls's and Scanlon's Kantian views, and by Kant himself. Some may deny morality need advance the human good (e.g., Ross and Prichard are often read this way). But this is not the real issue between teleological (or consequentialist) and deontological theories. Nor is it the question whether consequences of actions matter in assessments of what it is right to do. (As Rawls says, any view that denies that "would simply be irrational, crazy" (TJ, 30).) Instead, underlying the rift between teleological and deontological (or at least Kantian) views are different conceptions of practical reasoning. They approach differently the question: How are considerations of individuals' good to enter into moral reasoning, and how are we to assess and promote their well-being? Philosophical utilitarianism implies a particular response here: If (1) facts about well-being are the *only* fundamental moral facts, and we assume that (2) these facts can be defined independently of moral concepts and principles, then we are led to the idea that (3) morality has no intrinsic value of its own, but is simply an instrument (one we would be better off without if things were different) for advancing the human good. This consequence, when joined with the normative claim that (4) no one's well-being counts for more than anyone else's, leads to the primary thesis of normative utilitarianism that (5) in appraisals of what is right to do ultimately we should be guided by what maximizes the sum total of individual well-being (cf. CU, 109-10).

Scanlon especially wants to reject the idea that all moral questions are ultimately reducible to decisions about what conduct and institutions maximize an aggregate. Something very different is involved in right conduct, just institutions, and in moral reasoning even at the most abstract levels. What this is, the underlying idea of Scanlon's contractualism, is best approached by focusing on the role of moral requirements in social interaction.

Every social group has its system of norms, certain rules and practices which regulate the conduct of members of the group and which are seen as essential to continued social cooperation. If we concentrate solely on the role of these norms in guiding conduct and regulating interaction, we are led to the idea that the purpose of moral requirements is to maintain and promote individuals' well-being. But exclusive attention to the action-guiding aspects of moral norms can lead to disregard of an equally significant function: they also have a place in the public life of the group in that they provide a common basis for argument and agreement among members regarding one another's conduct and expectations. In the course of justification within any social group, moral norms serve as the primary source of *public reasons,* the considerations we appeal to in order to justify and assess critically one another's actions on the basis of generally accepted criteria. As such, moral norms provide the basic framework for the assessment of individual and institutional conduct, its justification, criticism, and praise and blame.[3]

We might look upon this public role of systems of moral norms[4] in providing a basis for public justification and agreement as derivative of its regulative role. That people are able to justify their conduct is normally necessary for social stability, a condition of individual well-being on any account of morality. An example of a purely instrumental approach to the public role of moral rules is the tendency of some utilitarians[5] to depict moral praise, blame, and criticism as actions just like any other, to be assessed wholly in terms of their usefulness to aggregate well-being. Here the role of rules in public justification collapses; what we publicly *say* to one another to assess conduct is to be assessed (presumably nonpublicly) in terms of the conduciveness of our pronouncements to overall well-being.

In stark contrast, contractualism assigns independent significance to morality's role in providing a framework for public justification, and seeks to account for the regulative purposes of moral norms through it. In answer to the question, "To what end are we to regulate our conduct?" the response is not "well-being," but "So that we may be able to justify our conduct to one another by considerations that everyone can reasonably accept."

Scanlon's most succinct characterization of contractualism is in these terms, in "an example" of "the kind of property which moral wrongness is":

> An act is wrong if its performance under the circumstances would be disallowed by any system of rules for the general regulation of behaviour which no one could reasonably reject as a basis for informed, unforced general agreement (CU, 110).[6]

---

[3] This says nothing about the validity of conventionally accepted moral norms, or their bases in self-deception or ideology. It simply provides a different approach to morality than utilitarianism or Hobbesian views. For further discussion of public reasons, see my "Reason and Agreement in Social Contract Views," *Philosophy and Public Affairs*, XIX, 2 (Spring 1990): 122–57.

[4] On the related idea of the social role of moral principles, see Rawls, "Kantian Constructivism in Moral Theory," The Dewey Lectures 1980, this JOURNAL, LXXVII, 9 (September 1980): 515–72, esp. p. 517.

[5] Henry Sidgwick, *Methods of Ethics*, 7th ed. (London: MacMillan, 1907), pp. 428–9; and J. J. C. Smart, "Freewill, Praise and Blame," *Mind*, LXX (1961): 291–306.

[6] More recently, Scanlon has characterized "a contractualist account of morality" as "a theory according to which an act is right if it would be required or allowed by principles which no one, suitably motivated, could reasonably reject as a basis for informed, unforced general agreement." "The Significance of Choice," *The Tanner Lectures on Human Value* (Salt Lake City: Utah UP, 1987), pp. 149–216, esp. p. 151. (Hereafter SC.)

Central to this definition is the idea that individuals are answerable to one another for their conduct, the claims they make, and the expectations they have regarding others' conduct and claims.[7] This is part of what it means to be reasonable. Answerability implies more than being held accountable (responsible) for one's acts. It means that we (or someone in our stead) be able publicly to justify our conduct, aims, and expectations on terms others could freely accept. This gives rise to the contractualist idea that the norms we rely on to justify as well as to regulate our actions and expectations are the product of an ideal agreement.

More needs to be said about the idea of reasonableness. But, first, consider the objection that there is nothing in Scanlon's definition, being reasonable, or even seeking a basis for public agreement, that would warrant representing morality as the *product* of a general agreement. Standards of reasonable acceptability and public agreement are *better* met by appeal to a prior realm of moral facts independent of us, which are evident to the light of informed reason (or to moral sense). If such facts exist, then rational and informed individuals would all, trivially, agree on them. They agree because all have an undistorted view of this independent moral realm. If so, there is no need to invoke a separate agreement as the *basis* for public agreement and what is reasonable; agreement would play no fundamental role in defining the subject matter of morality or the content of moral principles.

Scanlon says, "the appeal of contractualism . . . rests in part on a qualified skepticism" (CU, 118), by which I assume he means a rejection of an independent realm of moral facts and a special faculty of reasoning to ascertain them. This does not imply skepticism about the truth value of moral claims. Indeed, Scanlon's real point is that, if we see morality in terms of a hypothetical agreement among idealized individuals, then there is *no need* to invoke an antecedent realm of moral facts to provide a basis for moral truth and cognition. To clarify: imagine a community of free and informed agents, each of whom is sincere, conscientious, and motivated by a desire publicly to justify his actions, ends, and expectations to everyone else. Morality is the set of principles and rules which these persons could all affirm and to which they could jointly commit themselves, knowing that these terms (a) are to regulate their actions and expectations, and (b) are to serve as the ultimate public standards for justification, criticism, and the adjudication of conflicting claims.[8]

---

[7] Annette Baier suggested to me the centrality of answerability here.

[8] Although Scanlon does not explicitly define contractualism this way, I believe it incorporates the central features of his view—cf. CU, 111. The parallel with

As an account of morality's subject matter, Scanlon's view implies, not that our moral judgments are true when they match the judgments on which rational, informed individuals who have a privileged view of a prior moral order can agree. Instead, moral truth is defined by reference to the principles to which free and informed persons could reasonably agree under ideal conditions. There *are* no moral facts or properties prior to this hypothetical agreement. Significantly, this account provides a definition of moral truth and moral objectivity which does not require the ontological and epistemological commitments of rational intuitionism or its analogues. Contractualism thereby avoids the argument of J. L. Mackie's error theory (that there are no moral facts "in the world" for our moral judgments to be true of) (cf. CU, 108). We might then see contractualism as frustrating subjectivists' claims to be the spokesmen for the natural sciences. While it provides a conception of moral truth different from any that is appropriate for scientific claims, nothing in the account conflicts with descriptions of nature offered by natural science.

But contractualism must do more than this. Mackie's objection also might be avoided by a moral-sense view that is not committed to a realm of moral facts independent of human sensibilities. The main example here is ideal-observer theories, which define moral truth by the emotional responses of an ideally situated being with full knowledge and a perspicious view of the universe.[9] These accounts share in common with rational intuitionism the assumption that conditions of *observation* of some kind are fundamental to making correct moral judgments. This is the broader issue addressed by contractualism's focus on ideal judgment and agreement. It is not clear what the relevance of appeals to observation—no matter how ideal its conditions—are to grounding the truth of moral judgments.[10] In the context of moral justification, such appeals carry

---

Kant's third formulation of the categorical imperative, in terms of a "Kingdom of Ends," deserves separate discussion. Cf. TJ, 264.

[9] See Roderick Firth, "Ethical Absolutism and the Ideal Observer," *Philosophy and Phenomenological Research*, XII (March 1952). Others have lately relied on a similar idea for accounts of moral truth and objectivity. Gilbert Harman, "Moral Explanations of Natural Facts—Can Moral Claims be Tested against Moral Reality," *Southern Journal of Philosophy*, XXIV, Supplement (1986): 57–68; Peter Railton, "Moral Realism," *Philosophical Review*, XCV, 2 (April 1986): 190ff. David Lewis and Michael Smith, "Dispositional Theories of Value," in *Proceedings of the Aristotelian Society, Supplement*, 1989.

[10] Rawls raises this problem early on, in "Justice as Fairness," *Philosophical Review*, LXVII (April 1958): 164–94, esp. p. 194n; and it informs the account of objectivity in his "Kantian Constructivism in Moral Theory," pp. 554–72.

little argumentative force. Early in his essay, Scanlon says of both morals and mathematics that we need an epistemology appropriate to the subject matter of each, which explains "how it is possible to discover the facts about this subject matter *through something like the means we seem to use*" (CU, 104, emphasis added). Appeals to observation and a realm of moral facts prior to and independent of *reasoning* are not among the means we actually use in moral justification. In this context, observational appeals provide "reasons" of the wrong kind. To say there are moral qualities "out there in the world" or implicit in my perception of situations, which I detect by intuition or moral sense, does not provide a public justification of moral judgments, but is a way to cut short the process of justification. In moral argument, we seek to discover, not statements describing antecedent facts, but principled justifications of our moral convictions. This is the work that is done by appeal to ideal conditions for judgment and agreement.

This does not mean the notions of "moral fact," "intuition," etc., have no place in moral thought and argument. But the role they do have is different from the role of appeals to facts and experience in theoretical contexts. Appeals to moral facts, or to the observations of moral experience, are not appeals to brute data. They call upon the well-considered and firmly held moral convictions we share. And we share these convictions for a *reason*. The reasons that inform our intuitions, in moral contexts, are ultimately provided by principles.[11] Contractualism accounts for these principles by appealing to certain features of our practices of public justification. It explains morality by idealizing the ways of reasoning and justification we normally use in moral deliberation and argument. On this basis—and not by analogy with theoretical reasoning and its ideal conditions—contractualism constructs an "interpersonal standpoint" (CU, 117) that provides an objective point of view from which to ground claims about the validity of our moral judgments and conventions, and to assess arguments for competing moral principles. Our moral judgments are true when they accord with principles everyone could judge most reasonable from this perspective. Unlike rational intuitionism and ideal-observer theories, there is no assumption that moral facts or properties exist prior to and independently of *judgment* from this objective perspective. Rather, judgment from this standpoint specifies the realm of moral facts. Objectivity is, as it were, prior to the universe of moral "objects." The principles that ideally situated individuals could agree to define what the properties of rightness and wrongness are, and determine what natural and institutional facts are to be taken by us as morally relevant in our deliberations.[12]

---

[11] For example, what makes total domination of a person, or killing for enjoyment, morally relevant facts is that they violate principles we share. We say why these facts are morally relevant by appealing to these principles.

The controversy between moral realism and moral antirealism is a complicated subject. My intention is but to assign contractualism a place in this debate by indicating its distinct approach to these issues. Contractualism defines moral truth and objectivity, not in terms of antecedent properties (natural or non-natural), or ideal observational conditions, but by idealized conditions for judgment and agreement that model the nature and structure of our practical reasoning. (How it does this will be discussed in II below.)

Turn now to a third comparison, to clarify the role of the idea of *reasonableness*. One can see how, given contractualism's emphasis on public justification, this idea acts as a counterpoise to the utilitarian thesis that morality involves maximizing aggregate well-being. But reasonableness does more than alleviate the additive aspects of utilitarian and consequentialist views. Contrast contractualism with Gauthier's Hobbesian account. Free acceptability and mutual benefit are elements of any contract view. How these ideas are related marks a significant difference between contract theories. Scanlon sees them as distinct and independent notions; Gauthier does not. He represents rational beings as agreeing to certain moral constraints on the direct pursuit of their ends, solely for the purpose of more effectively maximizing their individual well-being. Free acceptance of these terms is one means among others to achieve this benefit. In this way, Gauthier's[13] contract theory is more akin to philosophical utilitarianism than to contractualism; it holds that facts about individual well-being are the only fundamental moral facts. This resemblance is reflected in the parallel ways Gauthier and utilitarians treat individuals' desires. Both take desires as given, and as having value on their own terms, without regard to their origin or what they are desires for. Moreover, both define rationality in maximizing terms. The difference is that Gauthier refuses to aggregate satisfactions. There lies the significance of his appeal to a general agreement. As a bargain, the purpose of Gauthier's agreement is not to provide public answerability for the claims individuals make in the name of their well-being. Self-interested utility maximizers have no concern for that, or for the adverse effects their claims have on the good of others.[14] They have no desire to be reasonable.

---

[12] Here I interpret Scanlon, CU, 118. Rawls makes a similar point in "Kantian Constructivism," pp. 564–5, 568–9; see also, TJ, 45, 516f.

[13] "Moral theory is essentially the theory of optimizing constraints on utility-maximization"—*Morals by Agreement* (New York: Oxford, 1986), p. 78.

For Scanlon, general agreement is not instrumental to well-being; "it is, in a more fundamental sense, what morality is about" (CU, 128). His agreement is not then represented as a bargain or compromise among essentially conflicting interests; the kind of agreement morality involves is different. Moreover, individual well-being is morally significant, "not because it is intrinsically valuable . . . but simply because an individual could reasonably reject" in moral argument to general agreement, principles "that gave his well-being no weight" (CU, 119). Contractualism's denial of the intrinsic value of well-being means desires are not taken as given in moral evaluation, and assigned importance simply on the basis of their strength. Contractualism assumes an account of individuals' legitimate interests—desires that are (in Kant's terms) *worthy* of satisfaction—which is defined by reference to the codeliberations of Scanlon's parties and the moral principles to which they agree. I shall discuss these features in the course of elaborating on Scanlon's account of moral motivation and its relation to practical reasoning.

II. CONTRACTUALISM AND MOTIVATION

For any philosophical account of morality to be convincing, it has to connect awareness of moral requirements with action. An account of the nature of morality should show why people are regularly motivated to do what they believe to be right. And a normative moral theory needs to show why people should be moved to act on its substantive requirements. Both problems raise questions as to "why anyone should care about" morality (CU, 105), as defined by an account of its subject matter, or by the principles of a moral theory.

To lay the background for Scanlon's account of moral motivation, we need distinguish two different ways to think of reasons as instrumental, relied on to support two accounts of rationality. We often speak of the reasons a person has by reference to his primary desires. On one philosophical account, which I shall call "Humean" or "desire-based," a necessary condition of *A*'s having a reason to do some act is that the act advance the satisfaction of a present desire *A* has. Most accounts of this sort allow that one may have reasons of which he is unaware. Moreover, a good deal of idealization may be required to say what one's current desires, and thus reasons, are

---

[14] Gauthier's parties' lack of concern for the reasonableness of their claims comes out in their initial bargaining demands in the social contract; each demands the maximum, the entire cooperative surplus—*ibid.*, pp. 141–50.

(e.g., what one would want consistently and after full consideration). But the basic assumption of Humean views is that, for each person A, there is a set of primary desires he has at any time t, which are given to him by his nature and circumstances, and which provide the necessary basis for his having any reasons for acting. Gauthier, Gilbert Harman, Bernard Williams, and others hold this view.

We also often think of reasons less subjectively. For example, some people (e.g., children) have reason to provide for their health, education, or future, though they have no concern for them now. This presumes people have interests independently of present desire. Based on this, "interest-based accounts" say reasons are at least partially defined by certain (objective) interests. This presupposes a notion of persons' rational good. Reasons are still instrumental, so long as they are definable by reference to what promotes one's antecedent interests. Philippa Foot has held a version of this view. She contends that prudential considerations give agents reasons, without respect to their current desires.

On the basis of both conceptions, it has been argued (by Foot, Williams, Harman, Gauthier, and others) that the kinds of reasons moral considerations provide are instrumental. One can have reason to act morally only if it satisfies a primary desire with moral content (e.g., a desire to be just, fair, honest) or at least advances some other desire or nonmoral interest. If so, the kinds of obligations, hence reasons, to which moral considerations give rise are, at best, "hypothetical." No sense can be made of Kant's idea that moral reasons are "categorical" or of a special kind.[15]

Now, Scanlon seems to deny that either of these accounts of reasons is adequate. But he does not deny that we can have desire- and interest-based reasons (CU, 105). The two motivational questions he raises are framed just in these terms. The first question is: What is "the nature of the reasons that morality does provide, at least to those who are *concerned* with it" (CU, 105-6)? The second asks whether having moral motives "is compatible with a person's good or whether it is, as Nietzsche argued, a psychological disaster for the person who has it" (CU, 106). My concern here is the first question.

What is Scanlon getting at? He is not trying to show the amoralist, the person with no moral concerns, that he has reason to do what is right whatever his desires may be. The Hobbist addresses this question. The problem Scanlon raises is different: How are we to account for the *fact* that moral considerations have such a grip on us?

---

[15] See Philippa Foot, *Virtues and Vices* (Berkeley: California UP, 1978), chs. 10–2.

Scanlon takes it for granted that moral concerns move the great majority of people; moreover, "they strike those who are moved by them as reasons of a special stringency and inescapability" (CU, 106). Given this, his questions are: (1) What desire do we *already* have which leads us, almost universally, to take so seriously our awareness of moral matters (as opposed to other facts)? (2) What is there about the object of this desire which warrants such special concern?

One problem with rational intuitionism is that it provides no answer to (2). Intuitionism holds that the recognition of certain moral facts (by one capable of intuiting them) is sufficient to give rise to a desire to do what is right *simply because* it is right. Moral motivation is defined by reference to a special cause, the intuitive grasp of non-natural properties or first principles.[16] Hume's argument against this view is based in the features of his account of human nature (reason is "inert" and so cannot give rise to any desire). One might contest that account (arguing that moral knowledge is sufficient to give rise to desire). But that would not remove the real problem, namely, since its moral psychology is so sparse, intuitionism gives us no reason to care about morality as much as we do. Knowledge is knowledge, and unless it is shown there is something special about morality, we have no more reason to be concerned with it than with knowledge of any other thing. Without further elaboration, our desire to do what is right appears arbitrary.

Here utilitarianism's account of morality's subject matter is thought to be particularly compelling: we are moved to act on moral requirements because they promote individuals' *well-being*. We are predisposed to commiserate with the pleasures, happiness, pains, and frustrations experienced by others, by way of the natural mechanism Hume called "sympathy" or by inculcated fellow feeling. In either case, it is a fact about us that we have concern for the good of others, and this is normally sufficient to account for our desire to do what is right according to common-sense moral requirements. Moreover, sympathy, once suitably educated and directed to aggregate well-being, can also provide sufficient motive to act on the principle of utility.[17]

The contractualist account of moral motivation differs from both of these views. Unlike intuitionism, the source of moral motivation is not a desire to do what is right just because it is right. It is a desire to

[16] See Sidgwick, pp. 23–5, 34–7, 39f.; W. D. Ross, *The Right and the Good* (New York: Oxford, 1930), pp. 157–60.

[17] See J. S. Mill, *Utilitarianism*, ch. 3.

act in ways that can be *justified* to others who are similarly motivated on grounds they could not reasonably reject (CU, 116).[18] This "desire for reasonable agreement," Scanlon contends, leads us now to be as concerned as we are about doing the right, fair, or just thing. Scanlon is pointing here to the fact that we do not want others to think, or even have grounds to think, that our conduct is unreasonable or our expectations are extravagant.[19] As Rawls has said, we want to live with others according to terms all would recognize as reasonable, from a perspective they would accept as reasonable. This means we are concerned that our acts and ends be justifiable from a public, and not just a personal, standpoint. This ideal of human relations is appealing to us, and we are drawn to affirm it. This is not to say we always act on this desire, or that we are not often mistaken about the right or most reasonable thing to do. Other sources of motivation compete with it (self-interest and so on), and we are subject to mistaken judgment and self-deception.

I take it that what is distinctive about the desire for reasonable agreement Scanlon assigns to us is this: unlike the intuitionist's desire to do what is right just because it is right, this desire presumes an account of the subject matter of morality as part of its content. In this way, it superficially resembles sympathy (construed as a desire for others' well-being). But unlike sympathy, the desire to justify one's actions also presupposes an account of justification (viz., on terms others cannot reasonably reject as a basis for unforced agreement). And that account of justification presupposes a notion of *principles*. The desire for reasonable agreement that accounts for moral motivation is a desire to act only in ways that can be justified to others according to principles they cannot reasonably reject, as a basis for general agreement. As such, it is what we might call a *principle-dependent desire*.[20] Sympathy is not like this; it is a desire for the well-being of particular individuals, or, once corrected, an abstract desire for the sum of well-being. Unlike the contractualist account of motivation, we can describe and understand the object of sympathy without appeal to any moral notions.

---

[18] In "The Significance of Choice" (p. 166), Scanlon defines "the basic moral motivation" as "a desire to regulate one's behavior according to standards that others could not reasonably reject insofar as they, too, were looking for a common set of practical principles. Morality, on this view, is what might be called a system of co-deliberation."

[19] Or as Scanlon says, "People are willing to go to considerable lengths, involving quite heavy sacrifices, in order to avoid admitting the unjustifiability of their actions and institutions" (CU, 117).

[20] Rawls describes the account of motivation in Kant and in "Justice as Fairness" this way. Cf. TJ, sect. 75. Kant defines "respect for the moral law" as "an interest in principles"—*Groundwork of the Metaphysics of Morals*, H. J. Paton, trans. & ed., (New York: Harper, 1964), Ak. 401n/69n, Ak. 413n/81n.

The significance of this fact will soon become apparent. First, I shall focus on certain features of the desire for reasonable agreement. It is clear Scanlon conceives of it as a primary desire. Like a desire for the good of others, the desire to justify oneself is something people normally want for its own sake. Contrast a Hobbesian contract view, like Gauthier's,[21] which says that moral motives, though real, are conditioned upon the expectation that one will benefit more than if he had no such motives and did not observe moral requirements. Second, as principle-dependent, the desire to act in ways that can be justified is a higher-order desire. It is not a desire for any particular object of action (like the well-being of others or oneself), but is a desire to *regulate* the objects of one's desires and their pursuit in ways (according to principles) none could reasonably reject. These two features imply that having an effective desire of this kind means that one will not (or at least will not want) (a) to impose certain claims and expectations upon others, even if one has a compelling desire to do so, and (b) to do certain things in pursuing one's ends, *because* making those claims and doing those things could not be justified to reasonable individuals.

What does this say about "the nature of the reasons morality provides . . . for those who are concerned with it"? First, it means we have reasons to act in noninstrumental ways, i.e., not for the sake of maximizing anything (satisfactions, well-being, etc.) or effectively promoting a particular end, but for the sake of regulative (moral) principles.[22] This does not mean that acting from reasons of this kind is aimless. Every action must have some kind of object for its end. The ends of action are normally given by the objects of first-order desires (e.g., pleasure, wealth, reputation, the good of another or one's countrymen). Moral desire specifies, not the specific end of action, but the *manner* in which one is to pursue specific ends. One with a higher-order desire to act in ways that can be justified does not want to take any means available (no matter how expedient, pleasurable, or likely) to meet his ends. When this desire is effective, we constrain the means we take to realize our ends, and regulate the aims that are the objects of our first-order desires. Different kinds of motives combine to determine our actions.

Second, the principle dependence of moral motivation implies that the *stringency* of the reasons morality provides is not determined by the strength or intensity of moral motives as compared with that of other desires. Instead, it is determined by the *content* of moral desire. To put this point in its context, it is a feature of Humean accounts that the relative weight of reasons is ultimately determined by the strength of the desires they promote. Where one's reasons conflict (as they inevitably will because of plurality of ends) the balance of reasons is decided, not by reference to these ends, the content of our desires, but by the relative intensity of the primary desires that provide a basis for practical reasoning. Underlying this conception is the basic premise that all desires are, for purposes of practical reasoning, *homogenous:* they are all to be put on a par, no matter what their origin or object, and assessed solely according to their relative strength (as experienced at the time of action, and after reflecting on all one's desires and ascertaining the likely consequences of acting on each).[23] This means that moral motives are to be assigned a weight just like any other desire, and are to be balanced off against them. So, if our moral sentiments are weak or waver under the circumstances, they should count for little in the process of deliberation, and provide reasons that are easily overridden by other concerns. On the Humean account, it would be irrational to assign moral motives, and the reasons they provide, any greater weight than their present strength allows.

---

[21] See *Morals by Agreement*, p. 169.

[22] It may be objected that acting on principles, or the rules they support, is instrumental to our being able to justify ourselves on terms others cannot reasonably reject. But the point is these same principles define the basis for justification and what count as reasons in moral argument. See below.

As an account of duty and obligation, or the weight to be assigned to moral considerations, the Humean account does not accurately represent our practical reasoning. Clearly we do not think that the stringency of moral duties (others' or our own) fluctuate with the current intensity of desire. Reference to desires seems out of place here. In our practical deliberations, we take for granted a background of moral norms. This background manifests itself when doubt arises regarding the legitimacy of our actions or ends: we do not normally ask whether abiding by these norms effectively promotes our prior purposes. Instead, we are concerned to establish whether our ends and proposed actions can be publicly justified to others according to generally acceptable moral criteria. By capturing this aspect of practical reasoning, contractualism (1) indicates what might be wrong with the Humean conception of the weight of moral considerations, and (2) provides an alternative conception of the process of reasoning about our own and others' duties.

(1) The force of the reasons morality provides is determined, not

[23] Homogeneity is implicit in views that identify rationality with utility maximization, in the completeness condition: that any two preferences are comparable. See *Morals by Agreement*, pp. 39ff. Scanlon's account of motivation, like Rawls's account of the sense of justice (TJ, 574–5), does not deny that such comparisons can be made, but holds they misrepresent "the nature of the reasons morality provides" (CU, 105).

by intensity of desire, but by the content of the desire for reasonable agreement, i.e., the principles that provide noninstrumental reasons. We have then at least one desire whose object demands that we not act on the overall balance of reasons (as decided by combined intensity of desire). The higher order desire for reasonable agreement is in effect a desire to *moderate* the balance of reasons and to limit first-order desires, as required by principles needed to justify our acts to reasonable persons. Given that content, it follows that acting solely on the balance of reasons settled by intensity of desire could never *satisfy* moral desire (because one's acts could not be justified, especially to those disadvantaged by them, by principles others could not reasonably reject). This means that moral motives, by virtue of their content, have the feature that the reasons they provide *override* the balance of reasons, as determined by the strength of all other desires.[24]

(2) How can this aspect of moral motivation be represented in an account of practical reasoning? From the Humean perspective, for certain reasons always to override all others implies that there are fixed desires whose intensity outweighs the combined intensity of all other desires. To assign to moral motives such an intensity would be highly unrealistic. There must be some other way to represent how moral desire is taken into account in practical reasoning, one that does not involve balancing the reasons it gives against others at all. This means that practical reasoning must have, at least for those with moral sentiments, a different *structure* than that represented by purely instrumental views.

Contractualism implies there are at least two different ways to deliberate about what we have reason to do.[25] These two different methods of reasoning are not exclusive, but supplemental, and one has priority over the other. The first form involves reflecting on one's ends as given by desires and (nonmoral) interests, ranking their priority and making them consistent, and then deciding the most appropriate means (effective, likely, inclusive, pleasurable, etc.) for realizing them. This form of deliberation is *agent-centered*, since it focuses on the present desires and interests of the self and asks what ultimately are my purposes, and what must I do to realize them. However we characterize this account of reasons (take any conception of instrumental rationality), moral sentiments (or interests, as the case may be) are not to be included in determining the balance of reasons. They must be taken into account separately, by a different form of deliberation.

---

[24] This has been challenged (by Foot, Williams, Susan Wolf, and others) on grounds that often prudential concerns or nonmoral ends seem justifiably to excuse us from moral duties. (E.g., it seems appropriate to break certain commitments to take advantage of an exceptional opportunity to further one's career.) But as we shall see, overridingness is a claim about the conclusiveness of the combined set of moral reasons in practical reasoning. There may be occasions where one may breach a specific moral rule to pursue matters of interest, but there must be other moral considerations that warrant such action under those circumstances. Contractualism clarifies when and why these circumstances arise: one may breach a commonly accepted moral duty when one's action under those circumstances accords with a rule that could be justified to others in a hypothetical process of codeliberation.

[25] I am indebted to Rawls for the idea of two stages of reasoning, which correspond to his distinction between reasonable and rational. Cf. Rawls, "Themes in Kant's Moral Philosophy," *Kant's Transcendental Deductions*, Eckhart Forster, ed. (Stanford: University Press, 1989), pp. 81–113, esp. pp. 82–8.

This second level of reasoning involves asking whether the kind and balance of reasons given by my agent-centered deliberations satisfy moral requirements, the object of moral desire. According to contractualism, this means asking whether my proposed actions, ends, and agent-centered reasons could be justified on grounds others could reasonably accept. This ultimately involves deciding what principles could be accepted by everyone in a process of ideal codeliberation. Considerations like this often guide us in our individual decisions about what we ought to do. Moral reasoning exercises, in effect, appellate review over our agent-centered deliberations. So construed, moral reasoning *subordinates* the balance of agent-centered reasons.[26]

Arguably, this kind of reasoning is also instrumental, since it is prompted by my moral motives, and involves taking the necessary means (deliberating and acting in a certain way) to satisfy the desire for reasonable agreement. But this is trivial. The important point is that moral deliberation, unlike agent-centered, does not involve reasoning according to principles or precepts that direct that I take the most expedient (i.e., effective, likely, pleasurable, efficient, etc.) means to given ends (e.g., the well-being of others or myself). Instead, it requires appraisal of these ends themselves (to decide if my expectations and claims are legitimate), and assessment of the instrumental means I have *already* decided are appropriate, according to a very different kind of principle: those which others could reasonably accept as a basis for public justification and general agreement.

---

[26] I do not contend that ordinary moral reasoning only involves reasoning about what others could reasonably accept. As Sidgwick says (p. 6), there are several "methods" implicit in ordinary practical reasoning. But this sort of reasoning occupies a significant position in our deliberations. I also do not contend that ordinary reasoning is structured exactly in the way represented in the text. Like the Humean account, the account of different stages of reasoning is an idealization.

What of the further objection, that there may be other higher-order motives to act on regulative principles which also purport to be overriding in the same way. Take Foot's example of principles of etiquette. Many people normally subject their pursuit of ends to these rules as well. How does the higher-order desire to be well-mannered or elegant differ from moral motives?

The content of moral motivation directs that *all* my acts and the content of all other desires be justifiable to everyone on terms they cannot reasonably reject. So it includes any other higher-order desire. This means that the reasons morality provides subordinate all other practical demands in deliberation. Rules of etiquette are not like this. One may breach these rules for moral and other reasons (e.g., to save another, or oneself, from destruction). These excusing conditions are part of the rules of etiquette. Morality also contains excusing conditions (accident, mistake, duress, lack of capacity, etc.). But moral excuses are justifiable only on morality's own terms, i.e., on condition they could be accepted by others as a basis for reasonable agreement. Arguably, rules of law are like this: one may not breach laws for nonlegal reasons, even moral reasons. But to sustain this convention, every legal system professes to be just and to fulfill all moral conditions, hence to be justifiable on terms all could reasonably accept.

So what distinguishes moral motivation from all other higher-order desires to act on regulative principles is that the set of reasons morality provides are *final* and *conclusive* in the context of practical reasoning. There are no higher standards, and no higher procedures, to which those with moral motives might appeal, to assess the system of moral considerations in the course of practical reasoning. Moral deliberation subordinates all other procedures. This is not a matter of choice, but is implicit in the nature of moral reasons. They override, not just instrumental considerations, but the demands of law, custom, and of social rules in general. In this sense, the desire for reasonable agreement is a *highest-order desire*.[27] It is in effect a desire to subject the reasons that result from all other processes of deliberation to the test of a final and conclusive procedure.

If we see the reasons morality provides as involving a different form of practical reasoning, one that assesses and subordinates the balance of instrumental reasons and the demands of all other forms of deliberation, then we have an idea how moral considerations offer reasons of a special kind. Those who are morally motivated have overriding reasons of a noninstrumental kind to act on moral requirements, *whatever else* their desires and interests may be. This does not mean moral principles are strictly "categorical." But contractualism indicates something of significance in Kant's notion.

---

[27] Cf. Rawls, "Kantian Constructivism," p. 533, on the sense of justice as a "highest-order desire." Suppose someone (e.g., Kierkegaard's aesthete) resolves never to breach rules of elegance; they override all other motives, even moral ones (e.g., he will not save a drowning child for fear of wetting his white suit). This person regards rules of etiquette as if they were moral rules. There is nothing about etiquette that would account for the special stringency he assigns to it. We can only account for it by the strength of this peculiar desire. Also, it is doubtful this person can even be ascribed the basic moral desire, that *all* his acts be reasonably justifiable.

III. THE REASON DEPENDENCE OF MORAL MOTIVATION

I consider now how contractualism can elucidate another feature of Kant's view, the idea that moral motivation is based in reason. Kant speaks of an "intellectual cause" of moral motives, and of reason "directly determining the will."[28] This is often construed as a causal claim about the origins of moral motivation: by rational deliberation alone, reasoning in accordance with certain principles, one without moral motives (the amoralist) can come to have an effective desire to act on moral requirements for their own sake. Whether this is a correct reading of Kant is not my concern. There is yet another way to construe the reason dependence of moral motivation.

I have argued, in effect, that having moral sentiments implies one's motivations have a certain *structure*, a hierarchically ordered complex of desires. Where does this structure come from? According to contractualism, practical deliberation involves: first, ascertaining the balance of agent-centered reasons and the actions needed to effectuate them; then, reflection on our proposed actions, desires, and agent-centered reasons, to see if they accord with principles that could be reasonably accepted by everyone in a process of ideal codeliberation. Moral motivation in turn is defined as a desire to act only in ways and according to reasons that could be justified by this hypothetical procedure. Having this highest-order desire implies that, whatever my agent-centered reasons, I have a willingness to subject them to a higher procedure, and a desire to act on its outcomes by regulating means and even modifying the ends that direct my agent-centered deliberations.

This allows the inference that the complex of motivations has a formal basis in practical reasoning. A hierarchy of motives (first-order . . . . highest-order desires) is imposed by the stages of reasoning we go through to decide what we ought to do. The structure of practical reason is, in effect, reflected in our desires. This structure thereby influences the content of our agent-centered desires—we often revise our ends if they do not accord with moral principles

---

[28] *Critique of Practical Reason*, Ak.79; also *Groundwork*, Ak.122n/Paton, 128n.

—and so contributes to the generation of agent-centered reasons and actions done for those reasons.

That moral deliberation imposes a formal structure on our motivations, thereby regulating our desires and actions, is a central feature of Kantian views.[29] (It is part of what Rawls means by "the priority of the right over the good.") The complex of motivations has this structure due to the mediating influence of moral desire. Moral desire itself is reason-dependent in a more direct way. According to contractualism, moral motivation is ultimately *principle-dependent;* it rests on a desire to act on principles that could be justified to reasonable individuals. To say this desire is principle-dependent means, first, we have this desire for a reason; moral motivation is not an arbitrary taste. It also means the reasons we have for moral motivations cannot be accounted for by prior and independent desires, like Humean sympathy or love of mankind. By contrast, contractualism implies that the reasons that support our moral motives are the same reasons that underlie our moral judgments. We can explain why we have these motives, and why we want to act on moral requirements, only by appealing to the content of our deliberations about moral issues, viz., our judgments about what acts could be justified by principles others could reasonably accept.

Now granted, we would not have moral motives were it not for moral education. We acquire moral motives over a long time, acting in our youth initially out of authority, often times fear of punishment, and later from a desire to please those we love or respect.[30] To say moral motives are reason-dependent does not imply a special causal process, through which the amoralist could acquire moral desires purely by intuition of or reflection on sound moral principles. But we must distinguish the conditions of the *origins* of moral motives in other desires, from the conditions needed to *sustain* our moral motives *once acquired.* Once one has reached maturity and his reasoning abilities are fully developed, what sustains moral motivation is not some other desire (like fear of punishment, or sympathy). It is the judgment and conviction that one's acts accord with principles: those which could be justified to reasonable individuals in a process of codeliberation. This desire is *reason-dependent,* because it depends on exercising our capacities for practical reasoning in a certain way. It requires an understanding of moral rules and an ability to apply them in judgment. And ultimately it involves an ability to reason about what actions and institutions could be justified to others by principles they could reasonably accept. So moral motivation is not like sympathy. It is not a desire or predisposition we naturally have which could be sustained *independently* of judgment and the exercise of our capacities for practical reasoning. Whatever role sympathy might have in accounting for the origins of moral motives, it is not what sustains them once they exist.

---

[29] More than a hierarchy of motives is required for a Kantian view. Harry Frankfurt [in "Freedom of the Will and the Concept of a Person," this JOURNAL, LXVIII, 1 (January 14, 1971): 5–20] says we have higher-order volitions to have certain desires and to be a specific kind of person, which regulate our first-order desires. But there is nothing specifically moral about these desires as Frankfurt describes them. They might as well be agent-centered (e.g., the desire to be elegant). Moreover, from where do these higher-order desires come? For Frankfurt, they ultimately stem from a kind of radical choice. Kantian accounts differ on both issues. More than being a desire to have certain desires, moral motivation is principle-dependent. And these principles, we shall see, are not based in radical choice, but in our capacities for practical reasoning.

[30] Cf. TJ, ch. 8, on development of the sense of justice; CU, 117.

A third way to describe the reason dependence of moral motives stems from our beliefs regarding the *status* of our moral convictions. For most of us, it is our belief that our moral judgments are sound which moves us to act on them. I would no longer want to abide by moral demands under all the conditions I do, did I not believe that my judgments and actions are supported by true or reasonable principles. But my belief that there are such principles is, according to contractualism, just my belief that there are principles that could be justified to others in an ideal process of codeliberation. If so, then my desire to act on these principles is sustained by the conviction that my moral deliberations are not in vain, but are answerable by this objective criterion, knowable to me through practical reasoning.

Arguably, it is misleading to say that belief that a judgment is true is reason-dependent. Instead, our beliefs are dependent (causally, let us say) on certain facts that obtain in the world. But consider again the contractualist account of moral truth. The "grounds of truth" are not any specifiable order of facts, natural or non-natural, which exist prior to and independently of human association and reasoning, and which are imposed upon us via observation or moral experience. Instead, it is the principles that *could* be agreed to in an ideal process of codeliberation. This is not a fact independent of human reasoning. It is a construction articulated by principles—an Idea of Reason in Kant's sense—based in the common reasoning of mankind, as we go about exercising our practical capacities to decide how to live together socially. We ultimately appeal to this construction to guide as well as to justify our judgments about what we ought to do. It also sustains moral motivation.

I have mentioned three ways moral motivation has a basis in reason, as opposed to having its basis in independently comprehensible natural desire (like sympathy): (1) the structure of our motivations reflects the structure of practical reason; (2) the content of moral motives themselves, of these highest-order desires, is supplied by ethical principles that govern our moral deliberations; (3) our acceptance that these principles are true, i.e., are justifiable as reasonable to others similarly motivated, maintains our moral motives. Still, is this sufficient for Kant's purposes, to define moral autonomy? Nothing has been said yet about the bases for ideal codeliberation and reasonable agreement on which we as individuals deliberate and want to act. Here we must look to the structure and details of Scanlon's hypothetical agreement. Scanlon says little about this in his initial article. He says it is unlikely that the principle of average utility could be justified as reasonable on a contractualist view. But suppose it could; then the sense in which moral motives have a basis in reason would be attentuated, since the ultimate principle practical reasoning provides would direct that we maximize average well-being. To avoid this problem (a heteronomous motivation on Kant's view) we need further specification of the ideal procedure of codeliberation.

In more recent work, Scanlon says that the moral principles contractualist agreement seeks are "for the [self] regulation of behavior," an idea which involves two components: a specifically moral capacity, "the ability to reason about what could be justified to others"; and a nonmoral "capacity for critically reflective, rational self-governance."[31] These are the basic capacities for practical reasoning. They enable individuals to formulate ends and to pursue them effectively, and to regulate their conduct within social groups. As such, these capacities provide the basis for persons' being the primary units of agency and responsibility, and their conceptions of themselves as free. Moreover, these capacities are conditions of our practical deliberations, both reasoning about our purposes, and understanding and acting on the results of the hypothetical codeliberations contractualism defines (cf. SC, 175). They are the practical powers that enable individuals to have moral motives, and to take an interest in, and act on and from, the codeliberation to moral principles which defines the subject matter of morality.

Suppose we construe this ideal codeliberation as a procedure that incorporates essential features of these capacities. This interprets contractualism as a version of Rawls's "Kantian Constructivism." According to constructivism: (1) first principles of right and justice, as statements about good reasons, are regarded as true or false by reference to the social role of these principles in providing a public basis for justification in an (ideal) society of reasonable and rational persons (a well-ordered society in Rawls, and the Realm of Ends in

---

[31] "The Significance of Choice," pp. 173–6.

Kant). Moreover, (2) these principles are specified by a procedure of construction, the form and structure of which express our free moral personality, and therewith essential features of our capacities for practical reasoning.[32] Scanlon's contractualism satisfies (1). But (1) alone does not, I believe, clearly ensure that the principles to which Scanlon's parties agree will not be based in considerations of individual well-being. If not, contractualism is just a more roundabout justification for utilitarianism. This presents a challenge to Scanlon's view. One way to avoid this problem is to structure the process of codeliberation so that a description of the two practical capacities enters into the reasoning by which moral principles are justified. Principles resulting from that procedure could then be construed as an *expression* of our practical conception of ourselves as free, and as reasonable and rational beings (cf. TJ, 252, 505).

If we see moral motivation in this context, we get a clearer sense of how moral motives might have their ultimate basis in practical reason. The principles that result from a constructivist procedure of codeliberation would be what we ultimately are trying to reflect on in our individual moral deliberations. So our desire to act only in ways that could be justified as reasonable can be construed as a desire to act on principles that are the result of a procedure designed to express our practical nature as free moral beings.[33] By acting on these principles and from these motives, we act for the sake of principles "given" to us by practical reason out of itself.

Kantian constructivism is a complicated subject I shall not discuss further. Whether Scanlon sees contractualism this way is a separate issue. My point is that a constructivist interpretation fits the primary aspects of contractualism, and provides a way to carry through one of Scanlon's primary aims: to show that the reasons morality provides, if not strictly categorical in Kant's sense, are at least of a special kind.

IV. CONCLUSION

To sum up sections II and III, contractualism indicates how morality can provide (a) conclusive reasons for acting, (b) of a noninstrumental kind, (c) which have a basis in practical reasoning. And it does this in a way that is compatible with the assumption that (moral) desire is a condition of having (agent-centered) reasons to act on moral requirements. But while it satisfies this condition, contractualism denies three basic claims of Humean accounts: (1) the weight of moral reasons is given by the intensity of moral (or other) desire; (2) the content of moral reasons (normative principles) is specified instrumentally, by what most effectively promotes prior and independent ends; and (3) reason imposes no substantive requirements on desire or action. To assume desire is a condition for having reasons does not require accepting these conditions.

---

[32] See Rawls, "Themes in Kant's Moral Philosophy," pp. 95–7.

[33] Cf. TJ, 572: "The desire to act justly and the desire to express our nature as free moral persons turn out to specify what is practically speaking the same desire." See also TJ, 256.

I have not argued that moral demands are "inescapable," or that moral reasons apply to everyone no matter what his desires and interests. Scanlon mentions the unavoidability of moral reasons, but does not address it (CU, 105). It raises the question whether there are what Thomas Nagel calls "objective (moral) reasons." One condition for the objectivity of moral reasons is that the soundness of moral principles be independent of the desires of actual individuals. Contractualism says the "grounds of truth" for moral judgment are the principles that could be agreed to by members of an ideal community, each of whom has a desire to find principles for regulating actions and institutions which will serve as a basis for public agreement. Whether there are principles meeting this description does not depend on the desires *we* now have, or even on our beliefs and conventions, and the agreements we might come to on these bases. It depends on the knowledge, deliberations, and motivations of these hypothetical persons. By abstracting from these features of our actual situation, the interpersonal perspective defined by contractualism provides an objective point of view. Our moral judgments are then true when they accord with principles that could be unanimously agreed to from this objective perspective.

This provides a necessary basis for objective reasons. Is it sufficient? Arguably, since morality affects not just belief but also the will, it must be shown how we can be motivated to act on moral demands. Here Humeans balk at the idea that morality provides reasons that apply to everyone. For how can awareness of moral requirements, even assuming they are true, move me to act in the absence of antecedent desire?

Even if we accept the Humean thesis that desire is a condition of the applicability of reasons, this condition can apply at most to agent-centered reasons. Moral reasons, by contractualist definition, cannot depend for their validity upon *our* antecedent desires. Instead, they explain *why* we have certain desires. Or as Nagel[34] says, they explain how moral motivation is possible. I have argued in effect that a constructivist contractualist account provides a way to approach that issue. What bearing it has on the broader unavoidability issue is not something I can discuss further here.

SAMUEL FREEMAN

University of Pennsylvania

---

[34] *The Possibility of Altruism* (New York: Oxford, 1970), pp. 12–5.

COMMENTS AND CRITICISM

THE CAUSAL HISTORY OF COMPUTATIONAL ACTIVITY: MAUDLIN AND OLYMPIA*

Tim Maudlin has, in a recent paper,[1] constructed a thought experiment that raises puzzling and important questions about what it is we attribute to a physical system when we describe it as performing a computation. His conclusion is that this thought experiment repudiates the possibility of a computationalist theory of consciousness. I shall argue that it would be more plausible to adopt instead the following conclusions: (1) the notion of "computing" is, like many other cognitive notions, best characterized in terms of a causal theory that asserts that computational activity is essentially the product of a certain causal history, and (2) the causal history essential to computation has a particular structure, one in which the computed input plays a peculiar role vis-à-vis the activity it induces in the computing system. The spirit of conclusion (2), moreover, is well applied to a wide variety of other cognitive activities, it seems to me, and thus illuminates something fundamental about the sort of causal interaction between a cognizing subject and a cognized object which is generally essential for many types of cognition.

Let us first introduce a conceptual tool that will be applied later. The tool consists of a distinction between two sorts of causes of physical events. Consider, e.g., the arrival of a letter in one's mail box. The causes of this event could be described in terms of, say, the various forces applied to the letter which induced the letter's motion from its point of origin to the mail box. I shall call such causes *active causes,* which work in general to cause events to occur to an object *X* by the net deployment of forces, or the transmission of energy, or the application of stimulants to or against *X*, perhaps by way of intermediate causal chains (supplied, e.g., by a medium through which active causes are transmitted). I shall distinguish active causes from *passive causes,* which work in general by the net nondeployment of forces (or energy or stimulants). A passive cause is nothing more than the nonoccurrence of an intervening event; a passive cause of the letter's arrival is that there was no insurmountable

---

* In the course of writing this note, I have profited from a number of discussions with John Winnie, though any errors in my analysis are certainly mine alone.

[1] "Computation and Consciousness," this JOURNAL, LXXXVI, 8 (August 1989): 407–32.

0022-362X/91/8806/304–16

barrier that deployed a force against the letter blocking the letter's motion. To say of some state of affairs that it works as a passive cause of an event that occurs to an object X is just to say that this state of affairs is such that it does not block or obstruct the active causes that induce the X-related event. The fact that the Granville police did not arrest me and throw me in jail this morning worked as a passive cause of my afternoon stroll to the Student Union, since this nonarrest allowed my stroll to transpire. This active/passive cause distinction is admittedly somewhat vague, but it is sufficiently precise for the purposes at hand.

I. MAUDLIN ON COMPUTATION AND GAA THEORIES OF COGNITION

What must a physical system do, and what must its structure be, if it is to be aptly described as a Turing machine running some algorithm π on some input τ? Such a system must contain, at least, an arbitrarily long tape and machinery that can manipulate the tape's contents. This machinery must, moreover, produce the correct output for a running of π on τ by following the "algorithmic rules" as laid down by π. A Turing machine's rule-following behavior consists of three elements: (1) the machine must write, at each step in the computation, either a 0 or a 1 at its current tape address, depending on the directives of π for the machine's current state and the original content of the current tape address. (2) The machine's read/write head must move, at each step, one address to the right or left on the tape, depending on π's stipulations, and (3) the machine must enter, at each step, the appropriate final state, again in accordance with π. To say of a physical system that it is currently running π on τ, then, is to say, first, that the system has been given τ as its input and that it is executing the appropriate sequence of write/move/change state operations (hereafter, W/M/CS operations) for an application of π to this input. Second, it must also be true that, had the system been given an input θ not equal to τ, the machine would have executed the appropriate W/M/CS sequence for θ. These two conditions combine to assert that the activity of the computational system is generally appropriate for all (actual and possible) inputs. I deem this latter condition the *generally appropriate activity condition for computation* (hereafter, the GAA condition for computation). Oversimplified, Maudlin's view takes the GAA condition for computation to describe what we attribute to a physical system when we describe it as computing π on τ.[2]

---

[2] Maudlin adds additional restrictions on computational systems to handle the problem posed by the "rock instantiation" (cf. p. 418), but this formulation of his theory of computation suffices for all present purposes.

It is interesting to note that GAA type conditions seem to work well to characterize a wide variety of cognitive activities. Consider, e.g., a GAA characterization for reading: to say of a subject *S* that it is reading a text *T* is to say, first, that *S*'s current verbal representation is (more or less) identical to the content of the text *T* he is currently presented with, and second, that had *S* been presented with some text *T′* not the same as *T*, *S*'s verbal representation would have (more or less) coincided with the content of *T′*. That is, to say of *S* that he is reading *T* is to assert that the architecture of the subject/text system is such that the verbal activity of the subject will generally be appropriate given the content of the given text, for the actual text *T* or any possible (readable) text. Or consider a GAA condition for seeing: to say of *S* that he is seeing object *O* is to say, first, that *S*'s current visual experience more or less coincides with the actual appearance of *O*, and, second, that, had some object *O′* not the same as *O* been present to *S*, *S* would have had a visual image more or less identical to the actual appearance of *O′*. The list of GAA conditions for cognitive activities could easily be extended to cover other cognitive activities, such as "referring," "knowing," "perceiving," etc.[3]

It is not surprising that GAA conditions seem to offer plausible analyses of cognitive activities. Such activities generally involve a subject (person or machine), who does the cognizing, and an object, which is somehow "cognized" in virtue of the subject's activity. (In the case of computation, the cognizing subject is just the computational machinery (minus the tape), and the cognized object is the input given initially in the tape contents.) It is frequently proclaimed that it is just in virtue of the fact that the subject's activity is causally attributable in some appropriate way to some object *O* that the subject's activity amounts to a cognitive act performed on *O* (and no other object). Now the generic notion of "causation" is often explicated in terms of the truth value of counterfactual statements, hence it has become rather standard to explicate the relevant causal connection between a cognizing subject and the cognized object in terms of counterfactual conditions asserting that, not only is the subject's activity appropriate for the actual object presented to the subject, but the subject's activity would also have been appropriate for some other object, had this other object been presented instead. This sort of condition is precisely what GAA conditions stipulate for cognitive activities in stipulating that the subject's activity be generally appropriate for any suitable object.

---

[3] I have no idea whether Maudlin himself would countenance the generalization of GAA characterizations to cognitive activities other than computation; I merely claim that such a generalization seems rather plausible, given certain simple intuitions about the nature of cognitive activity.

II. THE RIGID AND SILENT OLYMPIA

Maudlin's thought experiment reflects on the architecture and activity of Olympia, a candidate Turing machine. Olympia's tape consists of a sequence of water troughs, which individually record the data entry 0 or 1 by being either empty or full of water. One of Olympia's components is "Klara," an ordinary Turing machine programmed to run π. Another component is an armature, which is little more than a rod on wheels with some additional apparatus. If Klara were given τ as an input on the tape, she would perform a possibly long and complex sequence of W/M/CS operations before, in the end, producing the correct output for a running of π on τ. Now the armature is rigged with a water hose, and barriers, levers, and drain plugs are attached to some of the troughs, so that, as the armature runs on its wheels along the troughs, it mechanically empties and fills the troughs just as Klara would have had Klara run π on τ. Furthermore, by a clever trick of defining machine states "disjunctively," Maudlin ensures that the passage of the armature along the troughs amounts to Olympia's passage through the same sequence of machine states that Klara would have passed through had Klara performed this computation. In sum, the armature is rigged so as to mimic the W/M/CS operation sequence that Klara would perform in running π on τ. But the armature, unlike Klara, is merely executing a rigidly predetermined sequence of such operations. The only thing the armature can do is perform this unique sequence of operations. The activity of the armature is, in this sense, dumb. The armature's activity, clearly, cannot by itself do the work of running π on τ, despite the fact that it executes the appropriate sequence of W/M/CS operations for this computation. For given the armature's rigid structure, it is not true that the armature would have performed the appropriate sequence of W/M/CS operations had it been given any input θ not equal to τ. That is, the GAA condition for computation is not met by the armature's activity.

Maudlin's important insight is that the GAA condition can be met by supplementing the armature with machinery which essentially does nothing, i.e., which engages in no "activity" (according to a vague but common interpretation of this term) whatsoever when the input is τ. Maudlin imagines copies of Klara attached to floats; one such float is fixed in place in each trough. As the armature passes over each trough, the trough's float is released so as to "detect" the content of the trough. If the content of the trough should differ from what it will be if the initial input is τ, the trough's float will rise or fall, thereby stopping the armature's activity and causing the attached copy of Klara to be pulled into action, whereupon this copy of Klara will complete the W/M/CS operations appropriate for the given input just as an ordinary Turing machine would. If the content of the tape address is what it should be when the input is τ, the floats do not move at all, and hence the chains connecting the floats to the copies of Klara just lay slack, and the armature performs its work unhindered. The presence of these copies of Klara, along with the floats and chains, serves to support the required counterfactual condition: we can now say with confidence that, if any input θ not equal to τ were given on Olympia's tape, Olympia (construed now as the armature and copies of Klara working together) would execute the uniquely appropriate W/M/CS sequence for θ. But again, as long as the input is τ, no float moves, hence no copy of Klara is ever called into play, and thus this extra machinery does nothing. Olympia's entire activity on τ consists of the armature's operations as it runs along the troughs. Despite the intuition that this activity is "dumb," there now appears to be a serious case for the claim that Olympia succeeds in running π on τ just in virtue of this activity, along with the (inert) copies of Klara in the background, doing nothing but supporting the relevant counterfactuals. Undeniably, Olympia meets the GAA condition for computation, and surely, Maudlin plausibly claims, this is all that is required for a system to run π on τ.

If we grant that Olympia succeeds in performing this computation when given input τ, then it seems that we must also grant that Maudlin has succeeded in refuting a computationalist theory of consciousness. Such a theory, Maudlin conjectures, would claim that (A) it is sufficient for a system to be in some conscious state ψ that the system run some program π on some input τ. Also, such a theory would hold that (B) for a system to be in any conscious state, it must be describable as a nontrivial Turing machine running an algorithm on some input. Finally, any theory of consciousness will surely endorse the physicalist claim that (C) distinct systems that are engaged in identical sorts of activity will be in identical conscious states (if any). (C) is intended to ground the intuition that the presence or absence of any object that is causally disconnected from a system's activity cannot make a difference with respect to the system's status as conscious (or with respect to the quality of the system's conscious state, if it is conscious). Maudlin's purpose in constructing Olympia is to demonstrate the joint inconsistency of (A), (B), and (C). It follows from (A) that Olympia must be in conscious state ψ when she runs π on τ. But suppose we insert a "block" in between the gears of each of the Klaras' mechanisms such that these mechanisms are disabled and can no longer do the work of supporting the relevant counterfactuals. The activity of Olympia on τ (the armature's activity) is now a noncomputation, since the GAA condition for computation is no longer met. But the blocks can be inserted in a way so as not to interact causally with Olympia's activity in any way. They need not even be touching Klara's gears or engaged in any sort of interaction with any part of Olympia whatsoever. Given (C), then, the presence of the blocks cannot change Olympia's conscious state in any way: it must still be in conscious state ψ with the blocks in place if it was in state ψ without the blocks. But given (B), the blocked Olympia can no longer be conscious at all (and thus, of course, cannot be in state ψ) given the blocked Olympia's noncomputational status. The computationalist theory of consciousness that consists of the conjunction of (A), (B), and (C) is thus reduced to absurdity.

If it is true that Olympia is a legitimate Turing machine, and that she does indeed run π on τ in virtue of the fact that she grounds the GAA condition, it must nonetheless be granted that she is a decidedly unusual one. One of the most unusual facts about Olympia concerns the causal history of her activity when given the input τ. This causal history is such that Olympia will perform the correct W/M/CS sequence for this input even if her tape sensitivity is deleted, say, by removing the floats from the troughs and letting the armature make its run. This contrasts sharply with the causal history of Klara's (or any ordinary Turing machine's) activity when she runs π on τ, since if Klara were to lose her tape sensitivity (say, by deleting the reading component of her read/write head) she would almost certainly not proceed to execute the presumably long and complex W/M/CS sequence appropriate for a running of π on τ. I suggest that this difference between Olympia and Klara, the difference in the causal history of their "computational" behavior, bears close scrutiny.

The simplest method of rendering the Olympia/Klara distinction, it seems to me, is in terms of an active/passive distinction in the causal histories of their respective activities on τ. In short, the list of active causes of Klara's "π on τ" activity includes the informational content of the tape. More broadly, the "reading process" that gives Klara her tape sensitivity functions as an active cause of Klara's "π on τ" activity. This reading process consists, of course, of a sequence of "reading acts" which are individually performed at each step in the computation, and which enable Klara to react appropriately, given the content of her current tape address; however Klara's reading mechanism is structured, and however she executes this process, the following is almost certainly true:[4] the physical event or process that constitutes, in Klara, a reading of a 1 at a tape address cannot be the same type of physical event or process that constitutes a reading of a 0 at a tape address. If this were not true, Klara's reading mechanism could not distinguish between a 0-entry and a 1-entry, and thus Klara would lack the tape sensitivity required to run a computation. Putting the matter loosely, a 0-reading must consist of the transmission of a signal, presumably in the form of the application of a particular sort of force (or energy, or stimulant) from a tape address through the read head and into Klara's machinery, and a 1-reading must consist of the transmission of a distinct signal in the form of a sort of force (or energy, or stimulant) not identical to the 0-type force.[5] Klara's entire reading process consists, therefore, of a pattern of sequentially deployed forces exerted by the tape on Klara's computational machinery, a pattern of forces that contains the tape's informational content in coded form. These forces must surely be among the active causes of Klara's activity when she runs π on τ; without these forces, Klara lacks the needed impetus to execute the appropriate W/M/CS sequence.

In contrast, Olympia's reading process, and thus the content of her tape, is a merely passive cause of Olympia's "π on τ" activity. That is, the fact that the input τ is given on Olympia's tape simply results in the fact that the armature is "allowed" to perform its predetermined W/M/CS sequence. Since no floats move on this input, no copy of Klara is activated, and the armature's activity merely proceeds without intervention. Note, e.g., that Olympia's reading process on τ consists entirely of the troughs' chains simply laying slack, thereby conveying the information that the given input is τ. Hence the physical process that constitutes a 0-reading by Olympia is, amazingly enough, physically identical to the physical process that constitutes a 1-reading (both processes are constituted by a trough's chain remaining slack). But whereas in an ordinary Turing machine this fact would mean that the requisite tape sensitivity was simply missing, this fact does not, apparently, cause a problem for Olympia, given that she manages to execute the correct W/M/CS activity despite this fact. The reason for this is just that all the active causes sufficient for Olympia's "π on τ" activity are installed in the interior of Olympia's computational machinery; these active causes consist of the force that pulls the armature down the troughs, together with the forces exerted by the mechanical riggings (the hose, barriers, levers, and drainage plugs) that collectively perform Olympia's "π on τ" work. It is, I believe, the internality of these active forces, and the fact that these active forces do not trace down to the informational content of the tape through the tape reading process, that makes the claim that Olympia is computing the input given on the tape at least a little perplexing. But it is not at all obvious that the tape's content must function as an active cause of the Turing machine's W/M/CS activity in order for that activity to be legitimately construed as a computation of the input given on the tape. We investigate this issue in the subsequent section.

---

[4] The only exception would be if Klara herself had an Olympia-like architecture; here, I assume her architecture to be that of a garden variety Turing machine.

[5] Alternately, a 0-reading could consist of the nontransmission of any force, while a 1-reading could consist of a transmission of a nonzero force. While a 0-reading would thus amount to a kind of "passive" tape/machine interaction, it would nonetheless not be true that the entire reading process were passive, i.e., consisted of no force deployments against the read head. It is only reading processes which are entirely passive which I critique below.

III. DREAMER/BELL SYSTEMS AND COGNITION

Let us imagine the following scenario: a subject *S* has a piece of paper *T* before his eyes (or any ocular organ) on which the sentence *p* is written. *S*'s eyes are directed sequentially to each word of this sentence from left to right, and, as *S*'s eyes are pointed toward a particular word, *S* simultaneously pronounces this word to himself. Finally, let us add the assumption that *S* is asleep and that his verbal activity is merely the product of a purely auditory dream. It is simply a coincidence, in other words, that the content of *S*'s auditory dream happens to coincide with the content of *T*. Clearly, despite this coincidence, *S* is not reading *T*. This is because there is virtually no causal connection whatsoever between *T*'s content and *S*'s verbal activity, but the activity of "reading a text" surely requires some such causal connection. This is just to say, of course, that the GAA condition for reading is not met by this subject/text system.

Let us now "Olympia-ize" this example in the following way: in conjunction with our dreamer *S*, whose eyes trace the words in *T* while privately stating each word while his eyes point toward it, we rig up a "bell system." The bell system is designed so that, if the text before *S*'s eyes is just *T* (and thus contains just the sentence *p*), the bell does not ring. But if the text were to have been different, such that at any point the word *S*'s eyes point toward does not coincide with *S*'s inwardly pronounced word, the bell will ring at precisely this point, thereby waking *S* up. *S* will then proceed to read the remaining given text in the usual way, using the portion of his mind normally deployed by *S* when *S* is awake and reading. The presence of the bell system serves to guarantee that the GAA condition for reading is met by the dreamer/bell system: this system's verbal activity will generally be appropriate for the given text, whether the text is *T* or any other text, since, should the dream-induced verbal activity fail at any point to be appropriate for the text, the bell will ring, enabling the system to read in the normal way whatever text happens to be present.

Let us reflect now on *S*, his eyes studying *T*, dreaming his auditory dream, with the bell system in place. The bell will not ring, naturally, since the given text is *T*, and *S*'s verbal dream coincides precisely with *p*. The interesting, albeit odd, question we must now consider is whether the dreamer/bell system, conceived as a single unit, is reading *T*. The case for the claim that this system is reading *T* consists, apparently, entirely of the fact that it meets the GAA condition for reading. But I think there are rather compelling reasons for denying that this system is reading *T*, and thus for denying that the GAA condition for reading is in fact sufficient for reading. The principle problem with the claim that the dreamer/bell system is reading *T* concerns the causal history of the system's verbal activity. In particular, this causal history does not consist of the stimulation of the system's sensory organs by the text. (The system's sensory organs will include *S*'s sensory organs together with the bell system, itself clearly a sensory organ of some sort.) To the contrary, this sensory physiology is completely dormant throughout the course of the system's verbal activity. This is to say that the text does not function as an active cause of the dreamer/bell system's verbal activity. The presence of text *T* simply works as a passive cause of the verbal activity, since this presence does nothing relevant to the verbal activity but prevent the bell from ringing and thereby interrupting the dreaming process. The active causes of this activity will be the sort of causes that induce dreams, such as a brain-wave pattern characteristic of REM sleep, or perhaps the secretion of a chemical in the brain which is produced only in sleep and works to generate hallucination-like experiences. To say of any experience that it is a dream is just to say, it seems to me, that the experience is produced by just such a causal history, viz., a causal history that traces down to active causes entirely internal to the dreamer's own cognitive machinery.[6]

---

[6] This characterization is incomplete, clearly, in that it fails to distinguish dreams from other experiences whose active causes may be internal to the subject's own cognitive machinery, e.g., hallucinations or acts of the imagination.

Of course, the external world plays a role in the causal history of a dream, too, insofar as the world external to the dreamer's mind serves not to interfere with the ongoing dreaming process, say, by not waking the subject up. A causal history of a dream experience will locate the active causes of the experience in the interior of the dreamer's mind, and usually ascribe to the external world a purely passive causal role.[7] But this is the very type of causal history which generates the dreamer/bell system's verbal activity in our scenario, hence it seems quite clear that this activity is as much a dream vis-à-vis the dreamer/bell system as it is vis-à-vis the dreamer alone. The verbal activity of this system is not, I am suggesting, tied sufficiently tightly to the text to establish the activity as a reading of the text. It is, I claim, only in virtue of the fact that a text *T* functions as an active cause of a subject's verbal activity that it is correct to say of the subject's activity that it is a reading of *T*.

The point I am making concerns, in a clear way, the meaning of the term 'reading'. In short, I claim that "reading" is an activity a subject performs to a text with the use of the subject's sensory organs, i.e., by the active stimulation of these organs by a pattern of forces deployed, so to speak, by the text (usually by way of a medium like light rays). This pattern contains the content of the text in coded form, so that the subject can apprehend the text's content by interpreting the stimulated state of his sensory organs. One cannot read a text, I am claiming, when one's sensory organs are dormant and unstimulated any more than one can ride a bicycle when the muscles of one's limbs are dormant and unstimulated. The causally active role of the text in the causal history of the subject's verbal activity is simply essential for the subject to be reading the text; where this condition is not met, the subject is not reading the text even when the GAA condition for reading is met by this subject, and thus even when the subject's verbal activity is identical to what it would have been had the subject actually read the text. This is the moral of the dreamer/bell system for reading.

Dreamer/bell systems, it is interesting to note, are easy to construct for a wide variety of cognitive activities, and all such thought experiments seem to me to elicit the intuition that the system is not engaged in the sort of cognitive activity at issue. Consider a dreamer/bell system for "referring": our dreaming subject S sits before a table with a candle, and simultaneously dreams an image of a qualitatively identical candle. The bell system is rigged to ring only if the appearance of the table's actual contents fails to be qualitatively identical to S's dreamt visual image, so of course the bell does not ring. (Were the bell to ring, S would wake up and study the table's contents in the usual way; he would then attribute some property to whatever he observed on the table, and this reference to the table's contents would constitute an ordinary referential act.) S now performs, in his dream, the intensional act of attributing to the candle (the candle he is "observing") the property of being bright. The relevant question, of course, is whether the dreamer/bell system has hereby actually referred to the real candle on the table, given (1) that the system's phenomonological state (consisting of the visual experience of a candle and the attributive act) is identical to what it would have been had the dreamer been awake and thus referred to the real candle in an ordinary way, and (2) that, had the table's contents been different in any way, the system would have engaged in a phenomenological state that would count as an ordinary (perfectly legitimate) act of reference to whatever happened to be on the table. The most reasonable answer, surely, is that the system has failed to refer to the real candle, given the nature of the causal chain that induces the system's intentional act. The deviant nature of this chain, again, consists of the fact that the active causes of the referential act are intrinsic to the subject's dreaming process, and do not in any way trace down to the real candle. The real candle is merely a passive cause of the referential act, in that its presence simply prevents the bell from ringing and halting the dream-related active causes that generate the referential act. But the real candle's passive causal role is not of the sort that will connect the real candle to the referential act in a way that establishes this act as a reference to this candle. At most, the dreamer/bell system has referred to the dreamt candle, and not the real candle at all. The point here, obviously, is the same as for the reading scenario: it is necessary that an object O function as an active cause of a "referential act" in order for that act to be a reference to O. Similar thought experiments could be given for "perceiving," "seeing," "believing," "knowing," and many other cognitive activities. In general, for any or almost any cognitive activity, it is necessary that the cognized object play an active role in the causal history of the cognizer's activity for the cognizer to be actually engaged in "cognizing" the object; I hereby label this condition for any cognitive activity X as the *causally active*

---

[7] Sometimes the external world plays an active role in the quality of a dream's content, if externally supplied stimuli are "assimilated" into the dream content without interrupting the dreaming process (when, e.g., a subject dreams of being attacked by a tiger and wakes up with a cat scratch on his face). My point is made here by the example of dreams in which no such assimilation takes place.

object condition for X (hereafter, the CAO condition for *X*). When the CAO condition is not met by a subject/object system, there is no reason to describe the subject's activity as a cognitive operation performed on the object. In my view, it is surprising but not paradoxical that a subject/object system may meet the GAA condition for a cognitive activity without meeting the CAO condition for that activity. Such systems are instantiated by dreamer/bell systems of the sort considered above. The GAA and CAO conditions, I claim, are both necessary for cognition of any (or almost any) type. I suspect, moreover, that the CAO condition as I have cursorily stated it here could be refined so as to make the GAA and CAO conditions jointly sufficient for cognitive activity, though I do not take up the task of providing the needed modification here.[8]

IV. COMPUTATIONALIST CONCLUSIONS

It is by now clear that I believe there is a strong analogy between Olympia and a dreamer/bell system, and thus an interesting case to be made against Maudlin's claim that Olympia actually performs a computation on the input τ. Olympia, as noted above, undoubtedly meets the GAA condition for computation, but she does not meet the CAO condition for computation. The input τ does not function as an active cause of Olympia's W/M/CS activity when this input is given, hence there is no reason, it seems to me, to count this activity as a computation performed on the input τ. Olympia is engaged in an activity that is too "introverted," in a sense, to amount to a computation; since the active causes of her W/M/CS activity are supplied by Olympia's own cognitive machinery, her activity amounts to the machine analogue of a dream, or hallucination, or act of the imagination. Computation is, like virtually all cognitive activities, an activity that supervenes on a particular sort of causal interaction between a cognizing subject (here, the computational machinery) and a cognized object (here, an input given on a tape). Computation, moreover, is intrinsically a reciprocal, active causal interaction in that it requires that the machine actively transform the tape's contents (by writing operations) and that the tape's contents actively work to "transform" the machine's activity (its movements along the tape, its writing operations, and its state transitions) via the tape reading process. Olympia engages in the first, but not the second, form of active causal activity. But it is, again, only in virtue of the fact that the object actively induces the subject's activity that it is correct to describe the activity as a cognition (here, a computation) of the object. There is, in conclusion, no fact about Olympia's W/M/CS activity, or her architecture, that establishes her activity on τ as a computation performed on the input τ, despite the fact that this W/M/CS activity is the same activity that Olympia would have engaged in had she actually performed such a computation.

---

[8] It seems to me that what the CAO condition ought to say is not simply that the cognized object function as an active cause of the cognizing activity, but that the object do this in some particular way. I leave aside the problem of characterizing this way.

For all that has been said, perhaps one's critique of a computationalist theory of consciousness might resiliently insist that the GAA condition alone is sufficient for cognition, and thereby concede that, in the dreamer/bell/text scenario, the dreamer/bell system is, after all, reading the text (and, in the dreamer/bell/candle scenario, the dreamer/bell system actually refers to the real candle). My only response to this reply is that, for me, reflection on the dreamer/bell scenarios is sufficient to make such a proclamation sound somewhat dogmatic. I would be happy to avoid dogmatism altogether and simply speak of GAA cognition (which asserts that the GAA condition is sufficient for a cognitive act) as well as GAA/CAO cognition (which takes the GAA condition and the CAO condition as both necessary and sufficient under some suitable modification of the CAO condition, for cognition). Perhaps both senses of cognition are legitimate (though I will confess I doubt it).

Olympia, while she succeeds in performing a GAA computation, does not succeed in performing a GAA/CAO computation. Hence Maudlin's argument entails the following conclusion: the status of a system as conscious cannot supervene on a system simply in virtue of its GAA computational structure. There is nothing in his argument which counts against the possibility that a system is conscious simply in virtue of its GAA/CAO computational structure. For those of us who take 'GAA/CAO computation' as simply equivalent to the ordinary notion of "computation," this amounts to Maudlin's failure to refute a computationalist theory of consciousness.

ERIC BARNES

Denison University

BOOK REVIEWS

Innocence and Experience. STUART HAMPSHIRE. Cambridge: Harvard University Press, 1989. 195 p. Cloth $20.00.

Moral philosophy, according to Stuart Hampshire, takes "prevailing moral ideas" as the "raw data and starting-point" for a critical reconstruction of "those few well-argued conceptions of human powers which now actually influence people's lives and purposes" (179, 32). Hampshire draws his prevailing moral ideas from literature and history as well as from an extraordinarily wide-ranging and humane moral reflection. He distills from these ideas an account of human nature from which he draws normative conclusions about "the perennial topics of the underpinnings and origins of justice, of the universal and conventional elements in justice, and of the relation of private to public morality" (157). *Innocence and Experience* recapitulates several themes and arguments from Hampshire's earlier books, but it also contains much that is fresh. Although not all its arguments succeed, the book's thoughtfulness and intellectual vigor are remarkable, and it offers a systematic statement of what should be regarded as a leading position in moral and political philosophy.

I

The essence of humanity, Hampshire argues, is "a perpetual openness to new ways of life, to new forms of thought, to innovations in language and social arrangements" (30). This capacity for "imaginative invention" inevitably drives us to formulate diverse and irreconcilable conceptions of the good. In Hampshire's view, imagination ranks alongside intellect, which "unites humanity across all barriers" (30) just as imagination divides us. He emphatically rejects all topologies of the soul that place reason at the top of a hierarchy of faculties. The equal primacy of imagination and intellect as defining features of the human condition implies a "dual aspect" theory of morality (136). The first aspect is pluralism and sempiternal conflict among substantive visions of the good, since these are the products of imaginative invention and the centrifugal forces of linguistic diversity (43) and memory. (Hampshire argues that our memories are the source of "singularity" (114–5) and ultimately of imagination (127).) Joined with plural visions of the good, however, is a thin (137), universal (77, 140), transhistorical (142), species-wide (154) notion of minimum procedural justice consisting in the willingness to submit to "a regular and reasonable procedure of weighing

0022-362X/91/8806/317–24

claims and counter-claims" (63) through negotiation or arbitration or adjudication. Thus, Hampshire's prescription is this: "every population of intelligent persons should be a patchwork of minorities in the pursuit of deviant conceptions of the good, held together by a shared respect for fairness" (132). Strikingly, he concludes in favor of "Heraclitus's vision" that "ineliminable conflicts . . . and . . . rationally controlled hostilities . . . [are] not only normal, but also the best condition of mankind from the moral point of view" (189).

Just as pluralism of the good is a consequence of human imagination and diversity, minimum procedural justice "is rooted in" (169) the "species-wide and cross-cultural endowment" (119) of practical rationality within a single individual's deliberations (53, 169, 181, 187). Practical rationality Hampshire describes as "the practice of promoting and accepting arguments for and against a proposal" (53). Like Plato, Hampshire regards "governance of the soul and governance of the state" as mirrors of each other (182).

This is a dubious identification, however, for it overlooks a decisive disanalogy between the ways individuals and groups weigh arguments. Arguments in negotiation and adjudication are invariably strategic: the contending parties argue in order to prevail, not in order to achieve the fairest, best, or most informed outcome. In adjudication, the litigants employ rules of procedure tactically to exclude or conceal inconvenient evidence; in negotiation, the parties hide the weaknesses of their positions and disguise their bottom lines. Thomas Schelling[1] has pointed out that a bargain can be struck only when one party concedes, and a party concedes only because she believes that the other party will not. It follows that no bargain can be struck if the parties know what the others will actually accept; agreement requires bluffing. Bargaining and adjudication, even in good faith, are defined by opacity and cunning, not openness. As Hampshire characterizes it, practical reason is dialectic, in the sense of Socrates or Karl Popper: "a precise analogue of the requirement in a theoretical inquiry that all the accessible evidence should be assembled and assessed" (94). In adjudication, arbitration, and negotiation, by contrast, argument is rhetorical. While it may also be true, and interestingly so, that an individual deliberating about practical decisions often deceives herself and makes rhetorical appeals *in foro interno*—Hampshire suggests this in the book's final paragraph—the contending standpoints within a single person are nonetheless controlled by her overriding interest in making the best decision. This is almost never true in adjudication or negotiation.

---

[1] *The Strategy of Conflict* (Cambridge: Harvard, rev. ed., 1980), pp. 21–2.

Moreover, adjudication and negotiation are in important respects inconsistent with the openness to all claims which Hampshire promotes. The procedural rules of adjudication grant jurisdiction over some matters but (therefore) foreclose jurisdiction over others; the most important decisions of courts are often exercises of the "passive virtues," the studied refusal to hear issues on the merits because the court believes that a definitive ruling would be bad for the political process.[2] Similarly, the possibility of negotiation typically turns on excluding parties and issues from the table. Gag rules make procedural justice possible.[3]

Hampshire's derivation of minimum procedural justice from the essential features of human nature thus fails; whatever minimum procedural justice is "rooted in," it is not individual practical reasonableness. One should not assume, however, that this seriously damages the case for minimum procedural justice. Openness to argument and willingness to entertain seriously the claims and interests of others, including enemies and foreigners, are clearly appropriate candidates for universal moral requirements, particularly since violence and oppression seems to be the sole alternative. In Churchill's words, "Jaw-jaw is better than war-war." Nor is there anything minimal about minimum procedural justice: listening in good faith to the arguments of all interested parties is an ideal that no historical government has even remotely approached.

It is hard to see why procedural justice is the *only* universal requirement, however, as Hampshire himself comes close to acknowledging when he defines "great evils" as "those states of affairs which are to be avoided for reasons that are independent . . . of any specific conception of evil" (106). These include "[p]hysical suffering, starvation, imprisonment, the destruction of one's family or home," and he argues that these "have a clear priority in practical thinking" (106). This being so, it is somewhat puzzling that Hampshire stops short of the conclusion (pressed by Bernard Gert in *The Moral Rules* and *Morality*) that the requirement to refrain from inflicting great evils constitutes a minimum and universal *substantive* morality. One might object to this requirement that the infliction of the great evils may in some circumstances be justified; but this cannot be Hampshire's objection, for he insists that the violation of minimum *procedural* justice may in some circumstances be justified as well (140, 168, 171–2, 186). Universal duties are

[2] See Alexander Bickel, *The Least Dangerous Branch: The Supreme Court at the Bar of Politics* (Indianapolis: Bobbs-Merrill, 1962), pp. 111–98.

[3] See Stephen Holmes, "Gag Rules or the Politics of Omission," in Jon Elster and Rune Slagstad, eds., *Constitutionalism and Democracy* (New York: Cambridge, 1988), pp. 19–58.

"absolute," but only in the sense that they are not relativized to particular theories of the good, not in the sense that they can never be overridden (140). Moreover, the familiar fact that procedural rules inevitably affect, and often determine, substantive outcomes partially undermines the reliance Hampshire places upon the distinction between procedural and substantive justice.

His explication of minimum procedural justice is a bit perfunctory. He neglects the important differences between adjudication and arbitration, on the one hand, and negotiation, on the other. In Hampshire's own terms, adjudication—which yields all-or-nothing resolutions, winners and losers—is the procedural justice of innocence, whereas negotiation—always aimed at compromise and log rolling—is the procedural justice of experience. More troubling is his neglect of the culturally-relative character of procedural justice, the fact that adjudicatory systems proliferate in much the same way as theories of the good. (Indeed, on Hampshire's argument, they would have to. The machinery of procedural justice itself "has to be established by negotiation" (109), and thus must bootstrap itself into existence. This implies that systems of procedural justice are imaginative inventions, not intellectual deductions, and will therefore assume diverse forms.) The plurality of procedural justice systems weakens Hampshire's case for bifurcating pluralistic substantive morality from universal procedural justice.

II

Throughout the book, but principally in chapter four, Hampshire offers a highly original and compelling *catalogue raisonné* of human passions and powers. His discussion of the vast yet neglected role that reminiscence and nostalgia play in value formation is superb. Hampshire's scope is wide, and he has evidently written at a high level of inspiration; in the course of demarcating the places of imagination and intellect in human life he offers remarkable insights about such disparate topics as sexual love, language acquisition, nationalism, Nazi anti-Semitism, and philosophies of history. In breadth and depth, *Innocence and Experience* is reminiscent of Hannah Arendt's *The Human Condition* (without the Hellenism) and Roberto Unger's *Passion* (without the romanticism). All three books engage in the same project of drawing normative conclusions from a broad-gauge account of human nature—what Unger calls "the classical style of normative argument"—and it is noteworthy that they concur in the decisive role they assign to human inventiveness and openness.

The classical style of normative argument rests on the possibility of arguing from facts to values, and Hampshire devotes a large part of the book to vindicating this procedure by exorcising "Hume's ghost," which insists that the entire enterprise rests on a mistake. He responds that Hume erroneously conceives of moral evaluations as "commentary on, and criticism of, human performances" (102) rather than as steps in practical deliberation. This "displacement of agency into observation" (18) leads Hume to "the error of false isolation" (82) of value judgments from the contexts that alone give them their sense (83–8), and it is this error that leads Hume to deny the rationality of value judgments.

Hampshire supports these claims with two more detailed arguments. First, he advances a metaethical argument that, in its attributive uses, the word 'good' is not the name of a property but rather the adumbration of a "conditional wanting formula" (24) according to which "*x* is a good *S*" means roughly "if you want an *S*, *x* is what you are looking for." Thus, there is a systematic connection between the description of *x* as an *S* and the evaluation of *x* (95). Second, Hampshire returns to an argument he has made elsewhere, stressing that in practical reasoning we typically deliberate about singular statements of possibility: we wonder whether *X* would happen if we *Y*, or whether it would be morally wrong to *Z* (where *X*, *Y*, and *Z* are unique, nonrepeatable events). The former is "predictive" and the latter "moral," but Hampshire argues that the former is no more objective than the latter.[4]

It is unclear that Hampshire has succeeded in bridging the fact-value gap, however. Although there may be a "systematic connection" (95) between "*x* is an *S*" and "*x* is a good *S*," there is none between "*x* is a good *S*" and "*x* is (morally) good." (Let '*S*' equal 'Nazi'.) Thus, the first argument eliminates the gap between "*x* is an *S*" and "*x* is a good *S*" only to relocate it between "good" in its extramoral and moral uses. And the second argument eliminates the difference in objectivity between (singular subjunctive) predictions and (singular subjunctive) evaluations only to relocate it between singular subjunctives of either sort and scientific laws. In the end, this much of the gap remains: we always face the choice between accommodating human nature and resisting it, and morality can be read off of human nature only on the accommodation strategy. Aristotle's is an accommodation strategy, whereas ascetic religions adopt resistance strategies. Hampshire's conditional wanting analysis of "good," which rules out the claims (made by skeptics but also by mystics) that what is good may be wholly antithetical to human wants, implies an accommodationist commitment that cannot itself be derived from the facts of human nature without begging the question.

---

[4] Hampshire first ventured this argument in "Subjunctive Conditionals," *Analysis*, x (1950): 9–14, repr. in *Freedom of Mind* (Princeton: University Press, 1971).

III

Along with the key distinctions he draws between imagination and intellect and between thickly substantive and thinly procedural moralities, Hampshire is passionately concerned with the moral contrast between innocence and experience. "The idea of experience," Hampshire writes, "is the idea of guilty knowledge, of the expectation of unavoidable squalor and imperfection, of necessary disappointments and mixed results, of half success and half failure" (170). There is no question that Hampshire adopts the standpoint of experience, but his admiration for the virtues of innocence is genuine (172–7), and he insists that the standpoint of innocence—interestingly characterized as freedom "from the strains of a perpetual contrivance of means to ends" (149)—is the source of values essential to a tolerable life. Indeed, Hampshire speculates that nostalgia for lost innocence may be identical with our idea of pure happiness (149). Perhaps the most interesting and important moral problem Hampshire addresses (and it is noteworthy that few other contemporary philosophers even raise it) is how we can confront the truths of experience without falling into disillusion or cynicism.

These reflections inevitably direct Hampshire to the problem of dirty hands in politics. Hampshire, like Machiavelli, has little patience with the voice of innocence that requires politicians to abjure "deceit and guile, unjust violence and sudden aggression, ingratitude in relations with allies and friends" (163). Machiavelli argued that a prince cannot live up to Christian morality; Hampshire makes the problem harder still by asking whether the prince can respect even "the bleak negations of procedural justice" (164). His answer is yes, even though he insists that

> [a] political leader who allows a commitment to fair dealing and to substantial justice in all circumstances to override his determination to retain power, and to protect the interests of his country, will surely become ineffective and impotent; and such a person would generally have done better never to have sought or accepted public responsibilities (164–5).

His argument contains two points. First, "evil, in the form of the drive to domination, consists in the *uncompensated* violation of this basic justice" (186, emphasis added). But sometimes violations of procedural or substantive justice are "compensated" because the alternatives are worse. (And thus the violations are not evil? But Hampshire also refers to such decisions as striking a balance between evils (171). Thus he does not seem to have fully faced up to the problem of dirty hands, the paradox arising from the claim that in politics it may be right to do wrong.) Hampshire takes as his example the unjust victimization of a minority group in order to maintain the standard of living of a leader's constituents (186; also 171). Hampshire is prepared to concede that this injustice may under some circumstances be justified, for politicians have "obligations to protect the reasonable interests of their innocent fellow citizens" (170) even by ruthless means. This he calls the "Machiavellian thesis."

His second point is that determining the lesser evil requires the politician to weigh all the contending arguments, so that "the decision is not made by default" (171). Thus, in the weighing of evils and of interests, minimum procedural justice "plays the role of the scales" (186). More specifically, the leader must "establish the nature of the claims upon him in their own terms before he restates them in his own terms drawn from his own conception of the good" (187). Even ruthless and unjust political behavior must respect minimum procedural justice by hearing out the arguments.

It appears, however, that Hampshire has made the problem too easy, too "innocent," in two distinct ways. First, in real life, political expediency may well require the politician to violate not only substantive justice but even minimum procedural justice in order to govern effectively. (Hampshire himself seems to recognize this when he asserts (164–5) that expediency may be inconsistent not only with substantial justice but also with fair dealing.) Floor managers at political conventions labor to deny dissidents the opportunity to present their case; presidents appease their coalition members by refusing to meet with the members' adversaries; diplomats maneuver to shut inconvenient groups out of negotiations. All of them claim that these violations of minimum procedural justice are necessary for political success, and we have no reason to doubt them.

Second, the "Machiavellian thesis" speaks of obligations to protect constituents' interests, and seems to presume that only obligations of this sort will justify injustices. Thus, Hampshire's problem arises only when the politician "weighs two moral claims on him" (186). But the practicing politician confronts a meaner problem: he or she must commit injustices not only in order to fulfill moral obligations, but simply in order to acquire and retain power. Nor need the acquisition and retention of power be morally motivated; many politicians have no large ambitions toward the public good, but merely expect that they will muddle through their term in office as creditably as their opponents would. Thus, the *real* Machiavellian problem may be this: What moral sense are we to make of a world in which mediocre politicians must violate minimum substantive and even procedural justice simply to hang on to their jobs?

Despite these problems, *Innocence and Experience* is a wise and generous book that it would be difficult to praise too highly. Calmly and beautifully written, it manages to say important things about a wide range of human experiences within a very few pages. Notwithstanding his Blakean title, Hampshire is neither visionary nor romantic; his book seems more akin in spirit to Shakespeare, with whom he shares an affection for the frailties as well as the victories of the life we lead in a decidedly nonutopian world. *Innocence and Experience* may well fall into the select circle of books whose admirers will find their conversations beginning with the question, "Which are *your* favorite parts?" (For the record, mine are the sections "Nazism and Evil," "Attachment to the Past," and "Innocence and Experience Illustrated.")

DAVID LUBAN

University of Maryland

Metaphor: Its Cognitive Force and Linguistic Structure. EVA KITTAY. New York: Oxford University Press, 1987. 358 p. Cloth $59.00.

Eva Kittay argues that metaphors have a "second-order" metaphorical meaning in addition to their literal meaning, and that it is possible to provide a theory of interpretation for this second-order meaning. Thus, she argues against approaches such as Donald Davidson's,[1] in which it is argued that metaphors *mean* what they literally express, but that they may be *used* (for example) to call us to compare two objects or events.

Kittay's specific proposal is that the literal meaning of an utterance is to be fixed by the role that the utterance plays in some contextually determined *semantic field*—where a semantic field consists in the dependencies holding between concepts in a particu-

---

[1] "What Metaphors Mean," in *Inquiries into Truth and Interpretation* (New York: Oxford, 1984), pp. 245–64.

0022-362X/91/8806/324–30

lar domain. Metaphoric meaning (or second-order meaning) then arises from the introduction of relations from one domain or semantic field (the domain of the vehicle) onto another semantic field (the domain of the topic). The idea can be thought of as an elaboration of a suggestion due to Max Black[2] that metaphors introduce new relations into a "system of commonplaces."

My comments are divided into two sections. In section (I), I consider Kittay's specific proposal—in particular her thesis that metaphor consists in the transfer of relations from one semantic field to another. In section (II), I return to the more general issue of whether metaphor is a matter of use or interpretation, considering a number of arguments that Kittay offers in response to Davidson's "What Metaphors Mean."

I

To set up the discussion, an example of how Kittay applies her theory is in order. In Plato's analogy of Socrates as midwife, the semantic field of child delivery (according to Kittay) is structured as follows:[3]

main verb: help
agent: midwife
resulting state of affairs: verb: create (deliver)
    agent: mother
    result: child
instrument: potions and incantations

This structure is then imposed onto the semantic field of philosophical instruction in the following way:

main verb: help
agent: Socrates
resulting state of affairs: verb: create (deliver)
    agent: student
    result: ideas
instrument: dialectic

This is an interesting way to think about metaphor, and it gives some substance to Black's idea of transferring relations from one system of commonplaces onto another. The transferred relations turn out to be the grammatical relations of agent, patient, instrument, etc., which linguists call *thematic roles.*[4]

---

[2] "Metaphor," in Black, ed., *Models and Metaphors* (Ithaca: Cornell, 1962). Kittay, I should note, also takes issue with Black on a number of points; see 31, 181.

[3] This sort of analysis is also found in her joint work with Adrienne Lehrer. See Kittay and Lehrer, "Semantic Fields and the Structure of Metaphor," *Studies in Language*, v (1981): 31–63.

One important concern, however, is whether the above analysis is interesting because of semantic fields or because of the use of concepts like agent, patient, and instrument. Strictly speaking, these concepts may not be incompatible with the theory of semantic fields, but you do not need to believe in semantic fields to employ such concepts. If thematic roles are doing all the work, why do we need the semantic fields? In chapter 6 other kinds of semantic fields are introduced—in particular, contrast sets (e.g., the following is a contrast set, ⟨verbs of cooking: boiling, baking, frying, steaming, etc.⟩) —but thematic structure does most of the work in chapter 7 when semantic fields are applied to metaphor.

Kittay often speaks of the advantages of semantic fields and their holism over "atomistic" semantics, but cannot help making use of the resources of atomistic semantic theories at critical junctures. For example, in chapter 2 (59) she utilizes the notion of selectional restrictions from Katz-Fodor[5] semantics—which would have to be considered a paradigm example of an atomistic semantic theory. Elsewhere (58–64), she assigns interpretations to utterances *via* what looks like a compositional semantics.

This point is made more significant by the fact that semantic fields are given no real independent justification in the book. Kittay (35, 216), like a number of continental philosophers, uses an analogy to phonology and cites Ferdinand Saussure's "insight" that there are no phonetic primitives and that phonemes are defined by the role that they play in a system of oppositions. Strictly speaking, this is not an insight, but an empirical hypothesis, and the hypothesis has been in eclipse since the development of the theory of distinctive features by Roman Jakobson in the 1940s. A central point of modern generative phonology has been that there is a stock of primitive features common to all human language from which the phonology of a given language is derived. Modern phonology is in fact atomistic.[6]

---

[4] The theory of thematic roles has its roots in Charles Fillmore's case theory; see "The Case for Case," in E. Bach and R. T. Harms, eds., *Universals of Linguistic Theory* (New York: Holt, Rinehart, & Winston, 1968), pp. 1–90.

[5] See Jerry Katz and Jerry Fodor, "The Structure of a Semantic Theory," *Language,* XXXIX (1963): 170–210.

[6] Stephen Anderson [*Phonology in the Twentieth Century* (Chicago: University Press, 1985), pp. 118–9] puts Jakobson's perspective as follows:

> . . . Jakobson felt that the doctrine of the arbitrariness of the sign had to be subjected to important qualifications. While it is of course true that signs are arbitrary in that the link between a particular *signifié* and a particular *signifiant* is established by the conventions of a particular linguistic system, this

Independently of whether semantic fields are necessary, there are other concerns about the specific proposal. One such concern may be found in the following objection from Jonathan Adler. Suppose I am trying with difficulty to cut my steak and say, "this knife is a spoon." Either Kittay must claim that my utterance is not a metaphor, which seems false, or she must claim that I have introduced relations from one semantic field onto another. The problem with the latter solution is that knives and spoons intuitively belong to the same semantic field. If they do not belong to the same field in this case, then we may rightly ask for criteria of individuation for semantic fields. The danger of circularity should be obvious here. Kittay must avoid using the presence of metaphor as a criterion for individuating two distinct semantic fields in cases like the above.

There is perhaps another solution to this difficulty. There is really no need to say that the new relations have to come from *another* semantic field. One could very well say that the relations merely come from some semantic field or other (possibly the same one). In the knife and spoon case, it is possible to say that one is imposing the relations of a semantic field back onto itself, with the change being that the objects in the field are taken to participate in different relations.

Although not specifically related to the use of semantic fields, it is worth noting that Kittay takes the rather courageous view (64 ff.) that metaphor essentially involves some sort of semantic anomaly. The view is courageous, for it must account for apparently nonanomalous metaphors like 'Liddy is Nixon's plumber'. Kittay argues that the semantic anomaly of these constructions will become appar-

---

> does not imply that the sorts of things which can serve as potential *signifiants* are constrained only by the necessity to be different from one another. In particular, language has an essentially spoken character, as a result of which the *signifiants* are necessarily to be found in their logically primary form in the structuring of sound. Furthermore, not just any differences in sound are potentially relevant phonologically: a universal inventory can be given of the small set of distinctive features that function to differentiate *signifiants* in natural languages. Insofar as the *signifiant* of any sign must be made up of these same elements, the arbitrariness of the sign is considerably restricted.

Important work echoing the same perspective includes Roman Jakobson, C. Gunnar, M. Fant, and Morris Halle, *Preliminaries to Speech Analysis: The Distinctive Features and their Correlates*, MIT Acoustics Laboratory, Technical Report No. 13, 1952, repr. Cambridge: MIT, 1963; Noam Chomsky and Morris Halle, *The Sound Pattern of English* (New York: Harper & Row, 1968); and Halle, "On Distinctive Features and their Articulatory Implementation," *Natural Language and Linguistic Theory*, I (1983): 91–105. I am indebted to Ellen Broselow, Richard Larson, and Dan Finer for discussion of these matters.

ent in a fuller discourse—one in which we are told, for example, that Liddy is a lawyer. It is not clear that this move will succeed, however, for it seems to collapse the distinction between falsehood and semantic anomaly.

II

Kittay devotes a fair portion of her book to an attack on Davidson's theory of metaphor and his thesis that metaphor is part of the domain of language use, rather than meaning. Most of her criticisms seem to assume that Davidson's arguments are inextricably tied to his program in semantics, but he in fact offers several arguments in "What Metaphors Mean," which are independent of that program. We can call them the *argument from simile*, the *argument from dead metaphor*, and the *full paraphrase dilemma*.

The argument from simile goes roughly as follows (*op. cit.*, pp. 253–7). We are seldom tempted to invoke figurative or second meanings to account for similes, for it is clear in such cases that the author of the simile is inviting us to make a comparison between two things—and is not exhaustively specifying the properties shared by those things. Why should we not take the same stance toward metaphors? Kittay notes this argument, and her response is as follows.

> Simile accomplishes some of the same cognitive ends [as metaphor] but uses different linguistic resources. A simile is not a connotative semiotic and has no second-order meaning. But the nature of the comparison is metaphorical. In particular, the 'like' in simile should be understood metaphorically, for it signals a *cross-categorical* comparison (143, fn.).

Of course, this passage simply hands the point to Davidson, for it concedes that no second-order meaning is necessary to make metaphorical comparisons.

The argument from dead metaphor is simple enough. Davidson argues that, if there really is a special (let us say second-order) meaning, why does that meaning not become the literal meaning when the metaphor dies (*op. cit.*, pp. 253, 261)? For example, the phrase 'mouth of a bottle' presumably began as a metaphor, but whatever metaphorical meaning it once had is surely not to be found in the dead metaphor. Kittay, takes note of Davidson's argument, but seems to miss the point.

> Davidson suggests that if metaphors had a special meaning then 'we might be able to specify the special meaning of a word in a metaphorical setting by waiting until the metaphor die[d]' [Davidson, 1984, p. 253]. While it is true, as Davidson says, that this cannot be done, we ought not thence to conclude that metaphors have no special meaning. Rather, if instead of a second meaning, as in the case of ambiguity, they

> have a second-order meaning, we understand why this cannot be done. For a genuinely dead and unrevivable metaphor is one in which no second-order meaning is recognized by speakers of the language—the word acquires as a first-order meaning the sense which previously only contributed to its second-order meaning (142).

Davidson's point is that, on theories like Kittay's, the first-order meaning of the dead metaphor *should* be the same as the original second-order meaning, but of course it is not (as the mouth of the bottle example among others shows). And even if metaphors never die but merely lie dormant, the question arises as to why they never take on the second-order meaning while dormant.[7]

The full paraphrase dilemma is arguably the central point of Davidson's paper.

> There is, then, a tension in the usual view of metaphor. For on the one hand, the usual view wants to hold that a metaphor does something no plain prose can possibly do and, on the other hand, it wants to explain what a metaphor does by appealing to cognitive content—just the sort of thing plain prose is designed to express. As long as we are in this frame of mind, we must harbour the suspicion that it *can* be done, at least up to a point (*op. cit.*, p. 261).

Kittay must either admit that metaphor can do no more than plain prose, or give up the notion that there is a definitive cognitive content (i.e., a second-order meaning) attached to metaphor. She maintains that metaphors do not admit of paraphrase (thus rejecting the first horn of the dilemma), and argues that they cannot be paraphrased because of the nature of semantic fields.

[7] James Higginbotham has suggested that Davidson's argument assumes too much about what it means for a metaphor to die. Suppose we reconstruct Davidson's argument as follows.

> Suppose an expression *E* has a literal meaning *M* and is used metaphorically. By hypothesis, the metaphoric meaning is *M′*. Furthermore, if *E* "dies" we would expect *M′* to be its meaning. But when *E* dies it does not have the meaning *M′* but rather *M″* whose reference is that of *M′*. Conclusion: *E* never had the metaphoric meaning *M′*.

In the example above:

M: (literally) the mouth of the bottle
M′: what stands to a bottle as the mouth does to an animal
M″: the opening in the top of a bottle

The defender of metaphoric meaning might hold that there is no other way to get at *M″* other than through *M′*. Thus, the burden is shifted back onto Davidson to show how one gets at the referent of the metaphor without assuming the metaphoric meaning. I think Davidson can uphold this burden. Just as it is often argued that names and demonstratives need not refer via descriptions, it can be argued that dead metaphors refer to something without the help of metaphorical meaning to fix the referent.

> We can give an exposition of a metaphor but not a paraphrase—not if the metaphor is still a live metaphor and has not collapsed into literality. The irreducibility of metaphor is importantly tied to the incongruity between the domains of the topic and vehicle. That incongruity guarantees that a metaphorical predication cannot easily accommodate itself in the conceptual scheme which lies behind literal and conventional language (37).

One might ask why incongruity between two semantic fields "guarantees" that the relations between them cannot be characterized in literal language.[8] We are given no argument. But if Kittay is correct about this—that the incongruity of semantic fields prevents paraphrase—then she has new worries. Kittay is, after all, presenting a theory of metaphors, and is describing specific cases of metaphors in a book that is largely composed of literal prose. If there is more to a specific metaphor than she can describe in the language of her theory, that is just to say that her theory of second-order interpretation does not completely explain what the metaphor does. Moreover, to the extent that any theory of metaphoric interpretation will have to be couched in literal language, it will follow that there can be no theory that explains what the metaphor does. But this is just Davidson's point—that no theory of interpretation can explain what a metaphor does.

III

There are a number of interesting ideas in this book, and if the arguments do not always succeed, they at least advance the discussion. A number of the ideas can be divorced from the use of semantic fields, which, appearances aside, are not essential to Kittay's project for metaphor. If there is a problem with the book, it is that it underestimates the strength of Davidson's position. Kittay is not alone in this. Many philosophers working on metaphor are guilty of the same.

PETER LUDLOW

State University of New York/Stony Brook

---

[8] Kittay does not say metaphorical predication *cannot* be accommodated in literal language but merely that it cannot *easily* be accommodated in literal language. But if it is merely difficult to accommodate these predications in literal language, it would seem that one in fact *can* provide a paraphrase of the cognitive content of metaphors, or come as close as one likes.

Hilbert's Program: An Essay on Mathematical Instrumentalism. MICHAEL DETLEFSEN. Boston: Reidel, 1986. xiv, 186 p. Cloth $42.00.*

This clearly written book has two main goals: (a) to set forth an "Hilbertian" epistemology of mathematics, i.e., a theory of mathematical knowledge extracted from the writings of David Hilbert; (b) to show, contrary to the received wisdom, that Kurt Gödel's theorems have not discredited this mathematical epistemology.

Michael Detlefsen, through Hilbert, does to realism in mathematics what Bas Van Fraassen recently did to scientific realism: namely, just as Van Fraassen argues that scientific theories need not be, and anyhow cannot be shown to be, anything more than "empirically adequate," so does "Hilbert" argue that mathematics, to be good, need be only "real sound"—i.e., true to the so called "real" propositions of mathematics. The "real" propositions, are, roughly, arithmetic and algebraic (quantifier-free) identities. Since some propositions for Hilbert are, in fact, real, he occupies a middle ground between fictionalists like Hartry Field and realists like Gottlob Frege.

"Ideal" propositions and "ideal" proofs, corresponding to "theory" in Van Fraassen's scheme, are to be treated instrumentally: evaluated (as epistemic tools) rather than interpreted. This evaluation is done, for "Hilbert," in a theory called "metamathematics," which is itself supposed to be free of "ideal" proofs and propositions to prevent what Detlefsen calls (cf. 16ff., 62ff.) the epistemic "dilution" of mathematics. (The less squeamish Van Fraassen does his evaluations in model theory, which is applied set theory.) In fact, Hilbert suggested, a metamathematical consistency proof would be, not only necessary, but also sufficient, for the evaluation. His reasoning was: if mathematics is not "real sound" then it can "prove" a false identity; but any such false quantifier-free formula has, as a counterexample, a (provable, of course) true arithmetic identity. Thus, if mathematics is not real sound it is inconsistent. Similarly, a consistency proof allows one to replace any ideal proof of a *real* proposition P with a *real* proof of P: if P were false then mathematics would be inconsistent—contradicting the "real" consistency theorem. Thus, a consistency proof for Hilbert plays a role some-

---

* I am indebted to Haim Gaifman, Saul Kripke, Sharon Shelah, Charles Parsons, and Stewart Shapiro for extremely valuable comments and criticisms of my earlier versions.

0022-362X/91/8805/331–6

what like the existence of God in Descartes's *Meditations:* a guarantor of proved propositions.[1]

Now, just as Descartes did not intend to replace science by theology, Hilbert does not claim that we can or should replace mathematics by metamathematics. Rather, mathematicians continue doing what they have always done: discovering new theorems. They could hardly make these discoveries in metamathematics. The "ideal" formalisms of mathematics have other virtues than that of being, say, true—they are instruments suited to human inquiry.

As the Yiddish saying goes: if things are so good, why are they so bad? Why is it that the Hilbert program is considered "dead" or, at best, replaced by other research programs? Gödel's theorems, of course, especially the second. This purports to show that one cannot prove the consistency of any past or future actual mathematical theory within formal number theory. Presumably, formal number theory contains all the "real" proofs there are in mathematics.

Detlefsen, in a careful epistemological analysis, spells out this "standard strategy" for defeating Hilbert's program. I shall merely say that both the exposition of his position, and the analysis of the "Gödelian objection" to it are so clearly written that even the beginner in philosophy of mathematics could benefit from reading the book.

He has two methods of defeating the received wisdom, within the spirit of what he calls "Hilbertian instrumentalism." One is to note that, if "ideal" mathematics is to be taken seriously as instrument, then it must be suited to the finite capacities of human beings. The human brain cannot make use of all of the potential proofs in "ideal" mathematics. It is at least possible, then, that all the "Hilbertian" need do is prove the consistency of some subclass of all the possible proofs in mathematics—and there might well be a metamathematical proof that this subclass is consistent. This is the argument Detlefsen develops in chapter V. As Detlefsen points out, this is no argument for Hilbert, but only a refutation of the "Gödelian challenge." I think, moreover, that it would be difficult to designate the subclass of "ideal" proofs to be proved consistent, because we can always introduce new concepts and thus proofs by abbreviations, as is done in informal mathematics. Formally, the use of abbreviations makes it possible to know the existence of "ideal" proofs other than by actually exhibiting the proof in the basic language of the system.

Detlefsen's other "instrumentalistic" rejoinder to Gödel is to

[1] A major difference between Hilbert and Descartes, though, is that Descartes regards the absolutely infinite as clear, distinct, and real; Hilbert admits as "real" only propositions concerning the potentially infinite.

deny the received opinion that Gödel's second theorem demonstrates the impossibility of proving the consistency of "ideal" mathematics in "real" mathematics. Here his argument, as set forth in chapter IV, takes two prongs:

(a) Gödel shows that a certain number-theoretic sentence, which can be "interpreted" as the claim that "mathematics is consistent," cannot be proved—in number theory, and hence in "real" mathematics. He does not show that *every* sentence that expresses the consistency of mathematics is similarly unprovable. And although a number of writers have tried to generalize Gödel's theorem to show that every plausible candidate for the honor of expressing the consistency of mathematics is Gödel-vulnerable, Detlefsen, examining their arguments one by one, concludes that they fail to eliminate all possible candidates.

Detlefsen focuses his attention on the so-called "derivability conditions" which have figured in discussion of Gödel's theorems. Consistency of (a given system of) mathematics is, of course, the statement that $0 = 1$ is not provable (in the system). Suppose we want to represent the notion of "provability" (of some suitable system) by a formula of arithmetic, call it Provable(x). Now it turns out that Gödel's theorems apply to any formula that obeys the following three conditions:[2]

I) If X is a theorem of mathematics, and $[X]$ is (the numeral denoting) its Gödel number, then Provable($[X]$) is a theorem of arithmetic.

II) For every formula X of mathematics, it is a theorem of arithmetic that

If Provable($[X]$) then Provable([Provable($[X]$)])

This amounts to saying that condition I is itself formalizable in arithmetic.

III) For formulas X and Y of mathematics the following is provable in arithmetic:

If Provable($[X]$) and Provable($[X \rightarrow Y]$) then Provable($[Y]$)

This amounts to saying that the principle of modus ponens is formalizable in arithmetic.[3]

Mostowski and other writers have argued that any arithmetic formula that does not obey these three conditions cannot formalize the intuitive notion of mathematical provability. Since these three conditions are also sufficient to produce Gödel's theorem, their conclusion is that Gödel's theorem refutes Hilbert's program. Detlefsen argues, plausibly in my view, that no one has yet given a convincing reason that precisely these three conditions express the "essence" of provability. He therefore concludes that the case against Hilbert from Gödel's theorem falls short of demonstration.

---

[2] I am omitting discussion of the so-called "Diagonalization" property, which seems to me a property of the system, not of provability.

[3] These principles bear a formal resemblance to principles of modal logic, a resemblance noted by Kripke, Georg Boolos, and others, but I shall not discuss the significance of this ambiguity.

I think that most logicians will be unpersuaded by Detlefsen here (by 'most logicians' I mean Haim Gaifman and Saul Kripke); for I do not think that the derivability conditions are really the heart of the matter.

The reason that derivability conditions are needed to winnow out candidates is that, in arithmetic, there is *no* formula that *means* "provability." To say that a formula is provable in a system is to say that there is a finite sequence of steps leading from the axioms to the formula, and the *concept* of a finite sequence is not an arithmetic one. Naturally, we can "encode" every given finite sequence as a single number, just as we can encode each formula of mathematics as a single number, but this does not achieve a reduction of the concept of a finite sequence.

But there are a number of theories—such as set theory—in which one can express the concept of a finite sequence straightforwardly, as Frege showed. We might even argue that the set-theoretical definition of "finite sequence" has become canonical—in the sense that the epsilon-delta definition of continuity in analysis has become canonical. When a formalized notion has become canonical, it no longer makes sense to ask whether the formalized notion "expresses" the intuitive notion. If we consider an "applied" set theory with names for the primitive symbols of the language of mathematics, it is arguable that we know how to write down a formula that not only represents, but actually defines, what Hilbert meant by "provability"; and thus a sentence that paradigmatically expresses the consistency of mathematics. (In first-order number theory, we can encode the finite sequences using such artifices as the "Gödel beta function," but no such encoding can replace the intuitive second-order concept of a finite sequence, which is why we need derivability conditions.) Now Gödel's theorem shows at least that the consistency of set theory cannot be proved (even) in (all of) set theory,[4] and this result decisively refutes the Hilbert program as originally conceived.

The only way out of this for Detlefsen, it seems to me, is to deny that the definition of a proof in set theory is, in fact, "canonical." For the class of proofs is simply the intersection of all classes of formulas that contain the axioms and are closed under the rules of inference. This definition is, of course, impredicative, and there are mathematicians, such as Henri Poincare, who have argued that such definitions cannot replace the intuitions which inspire them.[5]

---

[4] ZF set theory with the addition of the names of the primitive symbols is obviously a conservative extension of ZF.

(b) A more interesting argument put forward by Detlefsen is, in effect, the following: even if the Gödel "translation" of "proof" (and thus "consistency") is accepted to be the only possible translations into arithmetic notation, what Gödel translated was the *realist* notion of proof, not the *instrumentalist* version.

Consider the following concept, which I shall call *Rosser provability*, in honor of its discoverer. Suppose we order all mathematical proofs, say by length, in a theory T extending number theory. Then, relative to this ordering, we can say: a formula is Rosser-provable, if it is provable, and no "shorter" proof refutes it. Let T' be the set of Rosser-provable propositions in the language of T. Of course, if mathematics is consistent, then a formula is provable if and only if it is Rosser-provable: so $T = T'$. On the other hand, even if T is inconsistent, not every formula will be Rosser-provable; and if T is inconsistent, therefore, T' will not be closed under modus ponens.

Now imagine that the "shorter" proofs are also more instrumentally useful than the longer ones. A Rosser-proved proposition, for the instrumentalist, would tend to survive the discovery of a contradiction in T. In other words, we would revise T in order to maintain existing Rosser proofs. Detlefsen speculates that there might be an ordering of the proofs of mathematics such that Rosser proof becomes a concept congenial to Hilbertian instrumentalism, though it has metamathematical properties quite different from the proof concept formalized by Gödel—in particular, Rosser provability violates the derivability conditions.

In fact, we can define "Rosser consistency": T is Rosser-consistent iff one cannot Rosser prove $0 = 1$ in T. But all this condition means is that, whenever the absurdity $0 = 1$ is provable in T, it is refuted earlier on. And thus it is trivial to prove the Rosser consistency of T (i.e., the consistency of T'), assuming that T is consistent. Indeed, it seems plausible that the refutation of $0 = 1$ should rank relatively early in any epistemically significant ordering of proofs, so that the Rosser consistency would become an empty concept.

Now I find myself somewhat sympathetic to the idea that "Rosser proof" is an idealization of some scientific behavior. For example, the mathematical foundations of some quantum field theories are such that nobody knows whether the mathematics is consistent or not. Nevertheless, physicists continue to make calculations, and the idea of "cordoning off" inconsistencies, i.e., refraining from using mathematical instruments where we know they do not work, seems to be applied daily.

---

[5] Cf. Parsons, "Frege's Theory of Number," reprinted in *Mathematics in Philosophy: Selected Essays* (Ithaca: Cornell, 1984).

Still, can "Rosser proofs" replace proofs *in Hilbert's program?* Note that such a replacement would make it, as a practical matter, impossible to prove anything, since one would have to check each and every previous proof to see whether the theorem had been refuted already. Also, even if we could set up an instrumentally useful ordering of proofs, I doubt whether it would correspond to some syntactic ordering (such as length).

But most of all, what has this to do with Hilbert's program? Recall that Hilbert sought to show that "ideal reasoning" would never lead to a false "real" theorem. But a Rosser-proved "real" proposition, in an inconsistent system, might very well be false, as Detlefsen recognizes (140 fn.): the counterexample might appear later in the ordering! Now if we know where the inconsistency lies, we might be able to construct an ordering of proofs where this phenomenon would not exist, or at least be minimized.[6] But in the interesting cases, of course, this does not come into question.

Now Detlefsen is aware of all this, and refrains from claiming that Rosser provability can be used to carry out Hilbert's program (140 fn.). His is the weaker goal of refuting the derivability conditions as a necessary condition on a formalization of mathematical proof. The mere possibility of an ordering relative to which Rosser provability becomes a model for a "strict instrumentalist" concept of proof, he claims, is enough to refute the derivability conditions.

I have already argued that the derivability conditions are not as crucial to the question of whether Gödel refuted Hilbert as Detlefsen and his opponents think. But even if defenders of Hilbert think it vital to refute these conditions, I am mystified why Detlefsen thinks that he can refute these conditions with a proof concept that he himself agrees is not likely to be acceptable to Hilbert! All the defenders of the conditions would have to claim is that they were attempting to capture necessary conditions on any proof concept that could be of use to the Hilbert program (not on any instrumentalistic concept of proof).

Irrespective of Hilbert's program, though, Detlefsen's discussion does suggest that Rosser provability might have its own merit as a model of actual mathematical practice in the physical sciences. Indeed, it might be more interesting if Detlefsen went on to develop his own instrumentalistic philosophy of mathematics than to continue to defend Hilbert's.

MARK STEINER

The Hebrew University / Jerusalem

---

[6] Thanks to Shelah for pointing this out.

THE JOURNAL OF PHILOSOPHY

VOLUME LXXXVIII, NO. 7, JULY 1991

REDUCTIVE AND NONREDUCTIVE SIMILE THEORIES OF METAPHOR*

METAPHOR is often defined as implicit or indirect comparison. Those who say this generally are committed to the possibility of reducing metaphor to simile (direct comparison) on the grounds that both are comparisons and with the belief that we can make better sense of simile than metaphor. According to this line of thought, our account of metaphor should be parasitic on our account of simile, since metaphor is just subtle simile. The argument, made explicit, has two main steps: reduce metaphor to simile, then conflate simile and literal comparison. 'Juliet is the sun' is nothing other than a shorter, neater, more powerful version of 'Juliet is like the sun', and to understand the former we must first understand the latter. The simile theory attempts to save literal semantics for metaphor by reducing metaphor to simile, and only succeeds if a conflation of simile and literal comparison is allowed.

Most simile theorists hold that on the surface 'Juliet is the sun' is obviously false, since Juliet is a young woman and one physical object cannot be simultaneously both a young woman and the sun. On the other hand, they point to the fact that it looks as if 'Juliet is like the sun' is trivially true. One advantage of simile over metaphor is that similes are usually if not always true.[1] (Donald Davidson, for example, claims that similes are always true because "everything is like everything else.") If we take metaphor to be a special case of simile, then we can see why a sentence that on most semantic analy-

---

* I would like to thank Robert B. Brandom, Joseph L. Camp, Jr., William G. Lycan, Stanley Munsat, Michael Resnik, Jay F. Rosenberg, Geoffrey Sayre McCord, and Josef Stern for helpful comments on an earlier draft of this paper.

[1] Robert Fogelin denies this claim, on the grounds that their simply sharing a property is not sufficient for saying that two objects are similar. See his *Figuratively Speaking* (New Haven: Yale, 1988), p. 59 ff.

0022-362X/91/8807/337-58

ses is false may nevertheless "ring true."[2] And we may do so without recourse to special metaphorical or figurative meanings.[3]

This view—the reductive simile theory of metaphor—underlies many current theories of metaphor and claims for itself several virtues. First, it resists appeal to multiple kinds of meaning, by relying solely on literal meaning (plus a few rules) to determine semantic content. Second, it holds out a greater promise for a systematic treatment of figurative language if figurative language can be explained in terms of the literal, since so many philosophers of language think literal talk can be explained compositionally. Third, the reductive simile theory makes sense of our attempts to understand specific metaphors. Upon encountering a new, somewhat obscure metaphor, say 'Juliet is the sun', we do not ask whether Juliet is or is not the sun but rather ask how Juliet is *like* the sun. Thus, the simile theorist claims, we intuitively reduce metaphor to simile every time we try to understand a particular metaphor. To find the meaning of a metaphor one finds the meaning of its corresponding simile. To find the meaning of the simile, one finds the meaning of the corresponding literal comparison.

Not all simile theories of metaphor are so explicitly reductive. Recently, Robert Fogelin has presented a defense of a nonreductive simile theory in his *Figuratively Speaking*. Fogelin's attempt to save the simile theory avoids the explicit reductionism that has characterized the theory in the past, but leaves us with the question of what use there is in assimilating metaphor to simile after all. After showing what is wrong with the standard simile theory of metaphor, which reduces metaphor to simile and simile to literal comparison, I shall address Fogelin's less thoroughly reductive position.

I. REDUCTIVE SIMILE THEORY

Although the details of the reductive simile theory may vary widely, certain claims define the theory. Claim (0) (below) is endorsed by most simile theories, but it is not strictly necessary.[4] The conjunction of (1), (2), and (3) is entailed by any reductive simile theory in virtue of being such a theory.

---

[2] Upon analysis, we shall see that rather than metaphor be a special case of simile, simile is a special case of metaphor. In *Languages of Art*, Nelson Goodman briefly remarks: "Instead of metaphor reducing to simile, simile reduces to metaphor; or rather, the difference between simile and metaphor is negligible"—(Indianapolis: Hackett, 1976), p. 77–8. I call simile a "special case" of metaphor rather than claim a reduction, because I do not want to go so far as Goodman's latter clause.

[3] In their adjectival forms ('metaphorical', 'figurative'), these terms are usually treated as equivalent, even though the nominal forms are not so treated. For example, 'metaphor' admits of both a generic interpretation, according to which it is equivalent to 'figure', and a specific interpretation, which denotes a particular sort of figure and no others. Since the adjectival forms are usually taken to denote the genus, I shall use them accordingly as interchangeable.

(0) Simile and metaphor play the same linguistic role: they make (or prompt us to make) comparisons.
(1) Any metaphor can be paraphrased as an identity or predication.
(2) For any metaphor of the form '*A* is *B*' or '*A* is a *K*', there is a function from that metaphor to a simile.
(3) The similes that result from (2) can be explained in purely literal terms.

In one form or another, this view is shared by such diverse writers as Cicero, Davidson, Paul Henle, Hegel, S. J. Brown, George Miller, Max Black, and Scott Buchanan, plus the myriad writers of textbooks on English composition and on rhetoric.[5]

The simile theory is interesting as a case of reduction in the philosophy of language, but it is also interesting because it underlies so many diverse theories of metaphor. Consider Black and Davidson, for example. Neither is usually found on a list of simile theorists. Black attempts to explain how metaphors get their special figurative meaning (really a sort of pragmatic meaning), while Davidson tries to show that metaphor has no special figurative meaning whatsoever.[6] Davidson even argues against several versions of the simile theory, and only emerges as a simile theorist at all once attention is paid to his view of the relation of semantics and pragmatics to metaphor. For Davidson, the literal semantic meaning of the simile is the basic pragmatic meaning of the metaphor. This is not so far from Black, after all, since Black claims that simile provides an adequate paraphrase of metaphor semantically but would deny that metaphor and simile play the same pragmatic role. Some writers, like Cicero, Buchanan, and Hegel, explicitly endorse the simile theory, without making explicit each step of the reduction. Many current writers on metaphor, however, explicitly deny the simple simile theory while implicitly holding a more sophisticated version, usually one that makes important use of the troublesome distinction between semantics and pragmatics.

---

[4] Designating the claim as '0' is meant to set it apart from (1)–(3), since it is usually (but not always) treated as an assumption.

[5] See Cicero, *De Oratore*, H. Rackham, trans. (Cambridge: Harvard, 1942); Davidson, "What Metaphors Mean," in Sheldon Sacks, ed., *On Metaphor* (Chicago: University Press, 1978), pp. 29–45; Henle, "Metaphor," in *Language, Thought, and Culture* (Ann Arbor: Michigan UP, 1965), pp. 173–95; Hegel, "The Metaphor, the Image, and the Comparison," in *The Philosophy of Art*, William M. Bryant, trans. (New York: Appleton, 1879), pp. 40–5; Brown, *The World of Imagery* (London: Kegan Paul, Trench, Trubner, 1927); Miller, "Images and Models, Similes and Metaphors," in A. Ortony, ed., *Metaphor and Thought* (New York: Cambridge, 1979), pp. 202–50; Black, "More About Metaphor," in *Metaphor and Thought*; Buchanan, *Poetry and Mathematics* (New York: John Day, 1929). Unlike the other writers noted, Black's endorsement of (0) is limited.

[6] Pragmatic meaning is roughly the interpretation generated after attending to both the strictly semantic (context-independent) meaning of the expression interpreted literally and aspects of the context of use which support or require a reinterpretation of the expression. Figurative meaning is one sort of pragmatic meaning.

My argument against the simile theory is quite simple. Simile theorists emphasize claim (2), usually assuming (1) and (3), with (0) in the background as the most basic assumption. My strategy is to show that neither (1) nor (3) is true, and that, without (1) and (3), (2) is of little help in developing an account of metaphor. With (1), (2), and (3) undermined, there will be no reason to grant (0) as an assumption. (There are independent reasons for rejecting (0), which will become apparent later.) Let us begin by noting one obvious problem with claim (2).

I.1 Claim (2).

> (2) For any metaphor of the form '*A* is *B*' or '*A* is a *K*', there is a function from that metaphor to a simile.

The simile theorist is quick to point out that generating a simile that corresponds to a particular metaphor is often straightforward.[7] From 'Juliet is the sun' we make 'Juliet is like the sun', from 'Richard is a lion' we make 'Richard is like a lion', from 'Old fools are babes again' we make 'Old fools are like babes again', and so on. For any metaphor of the form '*A* is *B*' or '*A* is a *K*' we first try the basic algorithm: let a 'like' follow the copula. This does seem straightforward. The simile gives us a sentence to analyze which is not obviously false and provides a structure that informs us that the meaning is comparative.[8]

---

[7] This claim is limited to those metaphors of the required form, and so is only of general use in a theory of metaphorical or figurative discourse if claim (1) is true.

[8] If we distinguish between the semantic role of an expression and its pragmatic role, we can see that adherents of claim (2) usually maintain either that

> 2.1.a. For any metaphor, there is a function from the semantic role of its corresponding simile to its own semantic role. (This function may or may not be identity.)

and

> 2.1.b. There is a function from the semantic role of the metaphor to the pragmatic role of the metaphor.

The simile does seem to be comparative; what is not clear is that the original metaphor is. Adding a semantic marker is not trivial. No one would argue that, since we can take any claim of the form '*A* is a *K*', add a 'not' after the copula, and still have a syntactically correct and semantically meaningful sentence, the deep grammar of all predications is negation. What is obvious in the case of adding 'not' is less obvious (but still true) in the case of adding 'like'. When interpreting literally, there is a big difference between saying '*A* is a *B*' and saying that '*A* is like a *B*'. The former entails *A*'s inclusion in the set of *B* things, while the latter does not. The 'like' weakens the semantic commitment of the original sentence.[9] Since the semantic commitments of the two claims differ, the addition of the 'like' is not trivial. Why should we think that it is trivial when interpreting figuratively? I shall come back to this question.

I.2 Claim (3).

> (3) The similes that result from (2) can be explained in purely literal terms.

To say what 'Juliet is the sun' means, according to the reductive simile theory, we must explain the respects in which Juliet is like the sun. The relevant meaning of the simile is an *interpreted* meaning; we must say what inferences the expression sanctions when read figuratively. Juliet is warm, she is bright and dazzling, she is the center of Romeo's world, his day begins with her, and so on. (Does she, like the sun, burn alive those men who draw too near?) Still, we are no nearer the literal, with its promise of ready-made semantics. 'Juliet is warm' is metaphorical in a way that 'The sun is warm' is not. (The vagueness of 'warm' need not concern us.) The same holds for 'Juliet is bright'. (It is silly to say "Get out your sunglasses; Juliet is bright today!") 'Romeo's world revolves around her' does not mean that the patch of earth upon which Romeo stands revolves around Juliet as the earth revolves around the sun; rather, it means

---

or that

> 2.2 There is a function from the semantic role of the simile to the pragmatic role of the metaphor.

Both 2.1 and 2.2 take us from the semantic role of the simile to the pragmatic role of the metaphor. 2.1 does this by way of the semantic role of the metaphor; 2.2 does not.

[9] The 'like' weakens the claim to which it is applied because the inferences that one may draw from *A*'s being a *B* are no longer justified by the claim that *A* is like a *B*. The weakening function of 'like' is discussed further in section II.3.

that Romeo places central importance on Juliet's happiness and well-being and has made her his raison d'être. *Explication of the simile leads us to more metaphors.* Not only that—the explication of any simile leads us to metaphors, and has to.

To see why explicating simile leads to metaphors, let us compare simile (either live or dead) to literal comparison. Recall that much of the appeal of the simile theory lies in its purported ability to convince us that there is no significant gap between the figurative and the literal, at least concerning metaphor. The deep structure of metaphor (the paradigm of figurative discourse) is simile, and simile can be understood wholly literally. Let us begin by addressing the naive-realist version of the literality thesis. Then, I shall show that some similes cannot be understood wholly literally (thereby undermining the universal claim of the simile theorist), and finally we shall see why no similes may be understood wholly literally.

First we must note that the simile theory is often accompanied by a naive-realist version of the literality thesis. According to the naive realist, in order to find the meaning of 'Juliet is (like) the sun', one finds Juliet, finds the sun, and compares them. The similes mentioned so far are all easily taken as having ordinary physical objects as referents of the *A* term and the *B* term. Not all similes are so cooperative. Although it is possible to consider the warmth and brilliance of the sun in trying to understand what it means to say Juliet is warm and brilliant (possible, although not very fruitful), what are we to examine when we try to understand T. S. Eliot's "He laughed like an irresponsible foetus"? Assuming, for the moment, that we know to whom the 'he' refers, how can we find an irresponsible foetus whose laughter we can study?[10] We cannot, since there is no such thing. Eliot augments the simile and gives clues for its interpretation by way of metaphor and further simile, saying

> His laughter was submarine and profound
> Like the old man of the sea's
> Hidden under coral islands
> Where worried bodies of drowned men drift
> down in the green silence,
> Dropping from fingers of surf.
>
> (from "Mr. Apollinax")

---

[10] This may seem to be a contrived objection, but see George A. Miller, "Images and Models, Similes and Metaphors," for a particularly explicit version of this view (in A. Ortony, ed., *Metaphor and Thought*, pp. 202–50. As John Searle points out (in "Metaphor"), most comparison theories are sufficiently Aristotelian to make such existential presupposition; such presupposition precludes an adequate account of similes containing nondenoting expressions.

Our task here is not to determine the exact meaning of the simile but to note that we cannot expect that for any simile we need merely to find the objects to which the terms refer and then compare them. A naive-realist version of the simile theory cannot handle all the similes we can understand, much less all the metaphors. In particular, it cannot handle similes containing nondenoting expressions.

Not all similes can be understood wholly literally. There is, as we have just seen, the problem of similes containing expressions that fail to refer (in the ordinary sense of 'refer'). A Dutch expression used to praise especially fine food may be translated as "It's like angels pissing on your tongue."[11] Although angel-lore omits reference to pissing (and one would expect that angels do not) and although there are not really any angels to do anything even like pissing, the simile does not say "This tastes like nothing doing what it cannot do." Waxing compositional does not help. Failure to refer is not the mark of the figurative, any more than falseness is. 'Angel' has both a literal and a figurative or metaphorical interpretation. If I tell my three-year-old niece that her dead dog is an angel now, I have used the term literally even though it fails to refer. If I tell her to be an angel and not interrupt, I have used the term figuratively and told her to be well-behaved, not to drop dead.

The fact is that no simile can be understood wholly literally. A sentence of the form '*A* is like *B*' is not ipso facto a simile. In 'A bicycle is like a tricycle except a bicycle has two wheels instead of three', 'a bicycle is like a tricycle' is not a simile. It is a straightforward literal comparison. I can justify my assertion of 'The engine of my boat is like the engine of my car' with a long tedious list of properties both have—in the same literal sense. (Both are physical objects, made of metal, plastic, and rubber, both power vehicles, both are of internal-combustion design, both run on gas, and so on.) Although of the correct syntactic form, such a comparison is not a simile. Neither the inferences it licenses nor the justification it requires involves figurative discourse in any way.[12]

---

[11] Thanks to Arthur Ripstein for introducing me to this expression.

[12] I am operating with an inferential role theory of meaning. I distinguish between three basic sorts of commitment associated with linguistic practice: identificatory, assertional, and expressive. We make commitments in referring, asserting, and invoking metaphors because we may be called upon to identify the referent of our terms, justify the assertional content of our claims, or make appropriate use of extensions to explicate metaphorically interpreted discourse. For a discussion of assertional commitment, see Robert B. Brandom, "Asserting," *Nous*, IV (November 1983): 637–50. For a discussion of asserting metaphors, see Merrie Bergmann, "Metaphorical Assertion," *The Philosophical Review*, XCI (April 1982): 229–46. For a fuller discussion of extended metaphor and expressive commitment, see my "Extending: The Structure of Metaphor," *Nous*, XXIII (March 1989): 17–34.

The widespread reliance on some sort of semantic marker (e.g., 'like', or 'as') to mark the distinction between simile and metaphor is possible because most theorists define simile as a comparison with a semantic marker, then define metaphor as simile with the marker left out, and then conflate the definitions. Explicitly to define simile as metaphor with a semantic marker and metaphor as simile without the marker would be circular, of course, but by failing to so much as wave their hands in the direction of a distinction between simile and literal comparison these theorists implicitly rely on such a circle. Once we recognize, however, that in moving from metaphor to simile we have not moved from trope to nontrope, from figurative to literal discourse, it is clear that the simile theorists have not completed their reductive program but, at best, relocated it. They still owe us a strict reduction from simile to literal comparison, but that is a reduction from simile to a class they do not and arguably cannot define.

Often those who try to define simile syntactically claim that for simile, unlike literal comparison, the *A* term and the *B* term must represent different classes. The simile is taken to be a comparison between two unlike objects, two objects in "essentially different categories,"[13] or two "distinctly different things."[14] Even assuming that we know whether two given terms refer to items in different classes, and most nonidentical terms do, this requirement will not distinguish between similes and literal comparisons. Some sentences (sentence types) of the correct form, with the cross-class requirement fulfilled, can be used either as a literal comparison or as a simile. Consider 'George is like the dog next door'. The sentence is a simile if it licenses inferences or requires justifications such as "Both bark but neither bites," but it is not a simile if the relevant response is "They both have brown hair, both weigh 220 pounds, both have big teeth and bad breath," and so on.[15]

The presence of a semantic marker may distinguish simile from metaphor if we are sure the expression in question is a trope, but it cannot distinguish simile from literal comparison. The cross-class requirement cannot distinguish simile from literal comparison, since even if we grant that the terms denote objects belonging to different classes ('human', 'dog') we can give the same sentence both a literal reading and a figurative reading. (And notice here that 'figurative' is just another way of saying 'metaphorical'.) Thus, the distinction between simile and literal comparison lies in the sort of interpretation an audience assigns (or ought to assign) to the expression.

---

[13] Raymond W. Barry and A. J. Wright, *Literary Terms: Definitions, Explanations and Examples* (San Francisco: Chandler, 1966).

[14] M. H. Abrams, "Figurative Language," *A Glossary of Literary Terms* (New York: Holt, Rhinehart & Winston, 1957), pp. 60–3.

[15] Actually, even this is overstated, since we may interpret 'barks' and 'bites' either literally or figuratively.

1.3 *Claim (1).*

(1) All metaphors can be reduced to or paraphrased as predicates or identities.

We have already discussed the process of generating a corresponding simile for a metaphor of the simple identity form. It is usually easy to apply the basic algorithm to sortal predication metaphors, such as 'Fran is a fox'. These are the only cases the reductive simile theorists discuss, and then not under these descriptions. Usually they take one or the other form as the paradigm for all metaphor.[16]

We have been talking as if all metaphors work the same way, but this has not been shown and is not a trivial assumption. *If* the various forms of metaphor cannot all be reduced to either predicates or identities (or to at least one of these two categories), then the claim (2) reduction briefly discussed earlier will be idle. Although it may be true that any assertion can be put into the form '*A* is *B*', it is not true that any metaphor can be. First, not every metaphor is an assertion. Second, and equally important, even if we make every metaphor into an assertion, we cannot always get the right terms into the *A* and *B* positions. Once these two points are established, we shall see that the simple transformation of predications and identities from metaphor to simile cannot carry the theoretical weight it has been assigned.

To see that not every metaphor is an assertion, one need simply look at logically complex metaphors, as in the following lines from Yeat's *Sailing to Byzantium:*

An aged man is but a paltry thing,
A tattered coat upon a stick, unless
Soul clap its hands and sing, and louder sing
For every tatter in its mortal dress,
Nor is there singing school but studying
Monuments of its own magnificence;
To the holy city of Byzantium.

In fact, this stanza is full of metaphors that are not assertions. (This is not quite the same as saying that they are not asserted.) Taking account of the logical form of the first few lines above, the following assertion can be attributed to Yeats:

---

[16] Even Fogelin does this; see *op. cit.*, p. 97.

> If the soul does not clap its hands and sing and louder sing for every tatter in its mortal dress, then an aged man is but a paltry thing, a tattered coat upon a stick.

Although we have an assertion, none of the metaphors is asserted since they are embedded in the conditional.

Next, to show that, even if one could make every metaphor into an assertion, one cannot always get the right terms into the *A* and *B* positions, we need to look at a variety of types of metaphor. There are at least six kinds of expressions which may be interpreted metaphorically, and it is not obvious that they can be collapsed into one. First, there are the *simple identities* of the form '*A* is *B*', where '*B*' is a name or definite description. This is the most frequently discussed metaphor form, and 'Juliet is the sun' is an instance of it. (Metaphors that make an indefinite description from a name, such as 'He is a Casanova', do not fit this category.) Next, there are *pure predications,* of the form '*A* is ϕ', such as 'Juliet is warm' and 'Juliet is brilliant'. Third, there are *sortal predications,* of the form '*A* is a *K*', which seem to combine the first two categories. Do we analyze 'Fran is a fox' as saying that there is a fox that is Fran or as saying that Fran has certain traits that generally characterize foxes? These first three types of metaphor do not pose obvious problems for the syntactic maneuvers required by the simile theorist. Similar to predications and identities are *substitution metaphors* achieved by substituting a term that does not literally apply for one that does, as in e.e. cummings's

> pity this busy monster, manunkind
> not

This sort of metaphor seems quite close to speaker reference, since it is clear that cummings uses 'monster' to refer to humanity.[17]

Most troublesome for the simile theory are metaphors that operate as functions. There are the *noun-function metaphors,* of the form 'The *B* of *A*', which apply to an object *A* a function *B* that does not usually have a value for that object. Examples include Blake's "The countless gold of a merry heart, The rubies and pearls of a loving eye," and sentences like 'Argentina is the England of Latin America'. It is not crucial that the article be definite: 'a cloak of silence' is also a noun-function metaphor. Finally, there are the *verb-function metaphors*. (These are perhaps the most common metaphors.) These are of the form 'The *A* *X*es', where *X* is something *B*s typically do. It is at least prima facie tenable to maintain that something like '*A* is *B*' is presupposed by such a metaphor. Examples of this form include Blake's "When the green woods laugh with the voice of joy," Eliot's "The river sweats/oil and tar," and Dickinson's "Frequently the hills are pink,/Frequently are brown./Frequently the hills undress/Behind my native town."

---

[17] Searle's account of metaphor is well-suited to substitution metaphors and any others that mimic speaker reference.

The corresponding similes for metaphors that are neither predications nor identities often involve quantificational terms. The basic algorithm is insufficient for generating a simile from Blake's "The rubies and pearls of a loving eye," but with some extra work we can get 'whatever a loving eye has that is like rubies and pearls'. The 'whatever' is quantificational and stays quantificational, because there is no object (however fictional, abstract, or linguistically constructed) to which we can attach a name and so generate a substitution instance of the quantified formula.[18] By making this transformation, we gain generality at the cost of accuracy and informativeness.

The same problem faces the verb-functional cases. Take, for example, one sportscaster's hyperbolic verb-function metaphor: 'When Sandy Koufax was young he could throw a strawberry through a locomotive'.[19] There are two ways to generate a simile from this sentence, and both involve the introduction of quantificational terms. The simile theorist might say:

(a) When Sandy Koufax was young he was like someone who could throw a strawberry through a locomotive.

or

(b) When Sandy Koufax was young he could do something like throw a strawberry through a locomotive.

(a) quantifies over people; (b) quantifies over actions. If we want our quantified simile to be close to the English metaphor, we shall pick (b). That is, unless we decide that every metaphor of the form 'The *A* *X*es' is to be understood as presupposing a sentence of the form '*A* is *B*' and that the sentence presupposed is the key to the meaning of the verb-functional metaphor. In that case, we would choose (a) and attempt to reduce this category to the category of simple identities. This second alternative explicitly follows the simile theory's reductionist strategy. The sportscaster, however, was not saying that Sandy Koufax was *like* someone who could throw a strawberry through a locomotive and that if we want to know what Koufax was like we should find such a person. We have already dismissed the naive-realist version of the theory. Treating the simile quantificationally leaves us with a permanent quantifier.

---

[18] Since we have already eliminated the naive-realist version of the simile theory, it is not important that our substituends be physical objects. On any account of quantification this quantifier remains.

[19] Thanks to Joe Camp for this example.

This brief look at a variety of types of metaphor has shown that generating an '*A* is *B*' sentence with the appropriate terms in the *A* and *B* positions is not always a trivial task. Most troublesome are the cases that leave us with ineliminable quantifiers in the "corresponding" simile.

A story about how we could determine the meaning of a metaphor can be told using the move to a quantificational simile, but a better story can be told without it. To return to the Koufax example, if the metaphor means that Sandy Koufax is like someone who can throw a strawberry through a locomotive, it does not really matter whether there is or is not such a someone. *What is important is what such a someone would be like, what traits he or she would have.* The problem with the simile approach to such cases is that it introduces unnecessary complications, sending us in the wrong direction, looking for properties shared by (in this case) Koufax and some nonexistent someone who really has the properties attributed to Koufax in the metaphor. We need to use the dummy subject to generate a simile, but to understand the metaphor we can and do skip the quantificational stage and ask, "What would Sandy Koufax have to be like in order to have such a capacity?" And we answer, he would have to have a frightfully fast arm and a keen eye. Such an answer provides a basis for understanding the metaphor.

The simile theorist will also have to use quantification in generating similes from metaphors involving pure predication. Given that the simile theorist is determined to make every metaphor into a comparison, 'Juliet is warm' will have to become 'Juliet is like a warm thing' or 'Juliet is like something warm'.[20] Unless one is from the San Fernando Valley, 'Juliet is like warm' will not do. The quantificational terms get the sentence into the right form, but they not only fail to add anything that can help us to understand the metaphor, they also weaken the sentence. Now we no longer attribute warmth to Juliet but rather something like warmth! Simile waters down the assertional commitment of a purely predicative metaphor.

---

[20] One might be tempted to deny that 'Juliet is warm' is a metaphor at all, since one of the dictionary's entries for 'warm' is 'cordial', which may be applied perfectly literally to Juliet. It is crucial to note that (a) not every dictionary entry represents a literal interpretation of the term being characterized, and (b) although 'Juliet is cordial' may most readily support literal interpretation, its purported synonymy with 'Juliet is warm' does not show that the latter should or must be interpreted literally in the context in question.

The reductive simile theory is untenable. It fails primarily because the reduction of simile to literal comparison fails. It also fails because the transformation of metaphor to simile does not always preserve meaning. In addition, we have seen that there are metaphors that cannot be paraphrased as an identity or predication (contra claim (1)), and this shows the limited value of the simile theorist's reliance on simply adding a 'like' after the copula. Further, the differences in interpretations of supposedly corresponding metaphors and similes which have been noted in this section give us reason to doubt that simile and metaphor serve the same linguistic purpose (contra claim (0)). Since both the reductive and the nonreductive simile theorist hold that the shared purpose of these tropes is to make or to prompt us to make comparisons, let us examine the nonreductive simile theory before explicitly addressing claim (0). After reviewing the nonreductive simile theory, we should be in a better position to see whether metaphor and simile play the same linguistic role.

II. NONREDUCTIVE SIMILE

A less reductive simile theory of metaphor has been presented recently by Fogelin in *Figuratively Speaking*. Fogelin's characterization of the simile theory is slightly different from mine, primarily because he sets out a view that accepts claims (0), (1), and (2) but rejects (3). Fogelin does not make the mistake of assimilating simile and literal comparison, and that is why I call his view *nonreductive*. Like other simile theorists, he does attempt to reduce metaphor to simile, and so the criticisms of (1) and (2) presented above do apply to his view. Most important to Fogelin is the claim I have identified as claim (0), which contends that simile and metaphor do the same work: they make or prompt us to make comparisons.

On Fogelin's account, the simile theory is characterized by two theses, which he accepts:

I. The *literal* meaning of a metaphor of the form 'A is a ϕ' is the same as the *literal* meaning of the counterpart simile of the form 'A is like a ϕ'.

II. The *figurative* meaning of a metaphor of the form 'A is a ϕ' is the same as the *figurative* meaning of the counterpart simile of the form 'A is like a ϕ' (*op. cit.*, p. 29).

Thus, according to Fogelin's view, given a metaphor of the form 'A is a ϕ', it literally and figuratively means that 'A is like a ϕ'. Although Fogelin thinks that the second thesis depends upon the first, in what follows I shall treat them separately. The first thesis is easily undermined, as we shall see, and it would be cheap and uninformative simply to dismiss the second thesis because the first fails. The problem with the first thesis does not directly undermine the second thesis, and so the second thesis should be addressed independently. As stated in *Figuratively Speaking,* Fogelin's position does not support the second thesis, but we shall see that attending to that position presents suggestions for the development of a more systematic treatment of figurative interpretation. Let us look first at how Fogelin distinguishes simile from literal comparison, for it shows how he characterizes the figurative, and then turn to what he says about simile and metaphor.

II.1 Simile and literal comparison. On Fogelin's view, the first main difference between simile and literal comparison is that simile arises from a special sort of interpretation on the part of both speaker and audience, and can be assimilated to indirect speech acts. In particular, he appropriates a Gricean speech-act framework. Although we shall see that this appeal to Grice is ultimately uncompelling, the fact that the same bit of discourse may be interpreted either literally or figuratively does lend support to the claim that the difference between simile and literal comparison is interpretive. The second main difference between simile and literal comparison, on Fogelin's view, is that their interpretations differ in how they rank the salience of the features of each of the items compared.[21] Fogelin claims that the salience rankings of the simile are the reverse of the salience rankings of its corresponding literal comparison. To see why this claim is too strong, let us look at one of Fogelin's examples.

Fogelin introduces 'Margaret Thatcher is a bulldozer' as an example of a sentence which most theories of metaphor would designate as false but which his designates as true. He claims that it is true because it is really a way of saying "Margaret Thatcher is *like* a bulldozer" and that sentence is true. The explicitly comparative statement is false if we treat it as a literal comparison since this means allowing "*the standard features* of bulldozers to fix the relevant feature space'' (*op. cit.*, p. 87, my emphasis). To do this would violate some key Gricean maxims, and that is our tip-off to treat it as a figurative comparison. When we treat it as a figurative comparison, we reverse the order of salience of features which is established on the literal interpretation.

---

[21] Note that this does not take into account the salience of the features of the context of utterance, which are crucial in taking a Gricean approach to the twice-true case discussed below. In Fogelin's examples, the figurative arises within the sentence, and the context of utterance drops out of the picture. This is one very serious flaw in his appeal to Grice.

Fogelin cites what he takes to be a highly salient point in bulldozer feature space, saying that Margaret Thatcher cannot ''move huge quantities of dirt in an efficient manner'' (*op. cit.*, p. 87). Since she cannot (really) move (real) dirt efficiently (by the standards set by heavy machinery?), uttering this statement flouts the Gricean maxim ''do not say what you believe to be false.'' Bear in mind that bulldozers move more than just dirt efficiently; they move rocks and fallen trees and construction debris and all sorts of things. A more generic description of what bulldozers do is that they push aside whatever is in their path (depending, of course, on the size of the bulldozer and the size of the obstruction). If we keep this more generic feature in mind, then saying that Margaret Thatcher is a bulldozer does not flout the Gricean maxim, for she is a strong woman, capable of quite literally pushing aside what comes in her path. (And metaphorically, she has been quite capable of pushing aside the liberal opposition.) To validate Fogelin's claim that this is a violation of the Gricean maxim, we must amend the maxim to state, ''Do not say what you believe to be *false on a literal interpretation*.'' Unfortunately, the revised maxim is much too restrictive.

Remember that Fogelin claims that in treating a comparison figuratively we reverse the order of feature salience established by a literal interpretation. On a literal interpretation, a highly salient feature of bulldozers is that they are machines. What Fogelin may have had in mind in saying that the order of feature salience is reversed in the move from literal to figurative interpretation is that, once we have called Margaret Thatcher a bulldozer, it does not follow that since bulldozers are machines that we can then call Margaret Thatcher a machine.[22] The inferences sanctioned by a literal interpretation do not follow. But of course, sometimes they do. There is nothing anomalous about saying ''Margaret Thatcher is a bulldozer. She's a big grey machine with the power to push aside whatever gets in her way.'' If we interpret 'machine' literally, we may agree with Fogelin, but if we interpret it figuratively, as we would if we were interpreting 'bulldozer' figuratively, then we must disagree with him.

---

[22] He explicitly offers a parallel argument in discussing 'Winston Churchill is a bulldog'.

The point is that figurative interpretation supports further figurative interpretation, not that we are restricted in what we go on to say when we are interpreting figuratively. If we mix the figurative and the literal in interpreting bulldozer features, then we get the anomalous results. The Margaret Thatcher example shows that we cannot find the distinction between simile and literal comparison in whether we allow the standard features of bulldozers to determine the relevant feature space, as Fogelin contends. This is because the expressions designating those very features may be interpreted either literally or figuratively and so cannot be used to distinguish the literal from the figurative.[23]

II.2 Thesis I (see above, p. 349). Any appeal to the literal meaning of a metaphor is problematic, unless one thinks that metaphors are syntactic types of some sort (and Fogelin does not think this). Fogelin very quickly admits that "in general, of course, sentences of the form 'A is a ϕ' do not have the same literal meaning as sentences of the form 'A is like a ϕ' " (*op. cit.*, p. 29). The problem with Fogelin's first thesis is that he wants to limit it to metaphors but he holds (correctly) that metaphors arise from interpretation and so he cannot really talk about a literally interpreted metaphor. Let me explain.

Apply the first claim to 'Fran is a fox'.[24] Literally interpreted, this means that Fran is a member of a certain class (foxes). The simile 'Fran is like a fox' is *compatible* with Fran's being a fox, but it does not assert that Fran is a fox. We have already seen that on a literal interpretation the 'like' weakens the claim to which it is added. What we interpreters confront is a sentence, a group of sentences, a text, a bit of discourse. Then we make metaphors in our interpreting. *Interpreted literally, there is no metaphor.* 'Fran is a fox' may be interpreted either literally or metaphorically, but it cannot be a literally interpreted metaphor. So, when combined with the view that metaphors are made by interpretation and not just found in texts, Fogelin's own qualifications that this conflation of the meaning of 'A is a ϕ' with the meaning of 'A is like a ϕ' does not hold with just *sentences* serves to undermine the first of the two theses that characterize the comparativist position.

---

[23] When these expressions are interpreted figuratively, and are so interpreted because the initial expression is figuratively interpreted, we then may call the subsequent (figuratively interpreted) expressions "extensions." Metaphors and similes are extendable; literal comparisons are not.

[24] Fogelin uses 'A is a ϕ' to include not only predications of properties but also all forms of metaphor (*op. cit.*, p. 29).

II.3 Thesis II (see above, p. 349). Fogelin takes the second thesis to depend upon the first. He thinks that "since metaphors and similes say the same things literally, their mismatch with context will give rise to the same figurative meaning" (*op. cit.*, pp. 30–1). The second thesis does not fall prey to the argument against the first, since metaphors and similes arise from figurative interpretation and so may be said to have figurative meanings (once interpreted). Since the first thesis is undermined, it would be best to take the second thesis independently. We must ask whether the figurative interpretation of a sentence of the form 'A is a ϕ' must be the same as the figurative interpretation of 'A is like a ϕ'. If so, thesis II would be correct.

If correct, thesis II would require: (A) that the semantic feature of 'like' that serves to weaken the assertional commitment of the literally interpreted claim does not operate on a figurative interpretation; (B) that the semantic feature of the 'like' that serves to signal that the claim is comparative when literally interpreted does function when the claim is interpreted figuratively; and, (C) that figurative interpretation cannot help but be comparative. (If true, (C) renders the 'like' eliminable.) The problem for the nonreductive simile theorist is to support (A), (B), and (C). The obvious way to do this is to support (A) and (B) independently, and then use them to support (C). Unfortunately, only (B) is true and then only in some cases. My strategy is to show that both (A) and (C) are false, and to let the simile theorist have (B).

Contrary to (A), 'like' as a weakener of the speaker's assertional commitment does function in figurative interpretation. To see this we must begin by examining the assertional commitments of a metaphor and then examine the assertional commitments of its corresponding simile. Shakespeare's Romeo does not simply say "Juliet is the sun," he says "Ah, what light through yonder window breaks, It is the East, and Juliet is the sun." Most metaphors are extended in their home contexts, and these contexts serve to limit the range of their possible extensions. Shakespeare's context serves to refine the metaphor's assertional commitments and to direct the audience to a limited range of endorsed extensions. In Shakespeare's play, when Romeo says "Juliet is the sun" it is *not* legitimate to extend the metaphor with "She's mostly made of gas." On the other hand, if we take the sentence out of Shakespeare's context and ask "Which extensions are permissible now?" we will find that the whole set of sun-related predicates becomes available. So, if all Romeo says is "Juliet is the sun," then Romeo is committed to licensing all the inferences one can draw from that: that she is 93 million miles away, that she is a celestial body, that she is mostly gas, and so on.[25] Obviously, Romeo has not specifically endorsed or asserted each of these extensions, but he also has not precluded their assertion by others based upon his metaphor. He has licensed the audience to use these extensions without specifically asserting or endorsing any.

Now we must consider the corresponding simile, and again, let us take it out of the Shakespearean context. If Romeo had said instead "Juliet is *like* the sun," and nothing else, he is not necessarily committed to any and all extensions of 'Juliet is like the sun' (as figuratively interpreted), though he is committed to some subset of the extensions. Which subset? We find out only by getting him to continue. If Benvolio were to say "Sure Juliet is like the sun—she's full of noxious gases." Romeo could justifiably reply, "I didn't say that she *is* the sun; I said she's *like* the sun. She is the light and warmth of my world." The disclaimer at the beginning does not preclude Romeo's choosing any particular extensions to endorse. It just signals to his audience that there are some that he will not endorse.[26]

These considerations show that when interpreting a simple unextended metaphor all the possible extensions are open to the audience, but when interpreting a simple unextended simile the use of 'like' suggests a more limited endorsement. In that case, all the extensions are open to the speaker, but not all are open to the audience, and until the simile gets extended, the audience cannot tell which extensions are unavailable. Thus, the 'like' in the simile does weaken the assertional commitment of the figuratively interpreted claim.

The other significant semantic feature of 'like' is that it signals that the sentence is a comparison. This feature must operate on a figurative interpretation if thesis II is correct. ((B) makes this claim for similes; (C) makes the claim for all figurative interpretations.) If we only look at simple identities and sortal predication metaphors, the addition of 'like' does seem to make the resulting sentence comparative. Again, we must ask whether the original metaphor was comparative, just because we can make a comparative claim out of it.

The problems that arise when we turn to substitution metaphors, noun-function metaphors, and verb-function metaphors suggest that (contrary to (C)) the original metaphor need not be seen as comparative at all. In discussing verb-function metaphors, we considered the hyperbolic 'When he was young Sandy Koufax could throw a strawberry through a locomotive'. There we saw that putting a 'like' into the original statement also requires adding an ineliminable quantificational term. We concluded that to understand the metaphor we do not go through these grammatical transformations but instead just ask, "What would Sandy Koufax have to be like for this to be true?" We can get rid of the 'like' altogether by asking instead how he would be, what traits he must have. We are not comparing Sandy Koufax to someone who can throw a strawberry through a locomotive; we are noting that he has a very fast arm and a very keen eye. So, it is not the case that figurative interpretation cannot help but be comparative.

---

[25] For more on this, see my "Extending: The Structure of Metaphor."

[26] In this respect the simile wrenched from context is similar to an extended metaphor within a context, except that the extended metaphor also directs the audience to a set of endorsed extensions while the simple simile does not.

With (A) and (C) undermined, thesis II is undermined, and with it goes the comparativist position as Fogelin has defined it. Although Fogelin takes thesis I and thesis II to define his position, we need not. Fogelin appeals to a simplified Gricean framework to establish an account of figurative interpretation, and in the next section we will examine the strengths and weaknesses of this appeal.

II.4 Figurative interpretation. Fogelin holds that figurativeness depends upon an incongruity between the utterance (interpreted literally) and its context of utterance. Fogelin's distinction between the literal and figurative requires the speaker's "*putting forward a claim figuratively rather than literally*" (*op. cit.*, p. 96) and the audience's interpreting that claim to get a good fit with context. The audience knows the speaker has put the claim forward figuratively if the speaker flouts Gricean conversational maxims. This may seem fine in some cases, but it is undermined by what Ted Cohen[27] has called "twice-true" metaphors.

Consider, for example, the case of the company president who looks out the boardroom window at a sky full of stratus clouds just before meeting with the company's union labor leader. The company president says to her assistant, "There's a storm brewing."[28] There is no incongruity between a literal interpretation of 'storm brewing' and this context, so on Fogelin's view a figurative interpretation would not be supported. But the figurative interpretation of

---

[27] "Notes on Metaphor," *Journal of Aesthetics and Art Criticism*, XXXIV (1976): 249–59.

[28] Notice the difficulty of generating the corresponding simile for this metaphor. 'It looks like there's a storm brewing' just seems to be a weakening of the assertional commitment and suggests a more literal interpretation. Note also that the literal interpretation is not necessarily comparative.

'storm brewing' is appropriate as well. Lack of fit with context is not the mark of the appropriateness of the figurative, nor is it necessary for the figurative.

We might try to strengthen Fogelin's position by pointing out that it is not simple lack of fit that he has in mind. He says that "with a figurative comparison the speaker flouts, or at least violates, standard [Gricean] conversational rules and thus engages the respondent in the task of making adjustments that will produce good fit" (*op. cit.*, p. 96). Basic to these maxims is adherence to the "cooperative principle," which enjoins us to "Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the exchange in which you are engaged."[29] So, the speaker violates, or at least seems to violate, a maxim, such as "Be relevant," "Do not say what you believe to be false," or "Avoid obscurity of expression," and the audience tries to assign a meaning to the utterance such that the speaker has lived up to the cooperative principle.

Let us go back to the twice-true case. Notice that a lot depends on what we take the purpose of the exchange to be, and that in this case, as in most, there is more than one possibility. On the one hand, if we take the company president to be talking about meteorology (perhaps expressing a concern that the weather will make a delayed departure from the office inconvenient), then her gazing at the stratus clouds outside is an important part of the context and the presence of those clouds will help to support the truth of what she says. On the other hand, if we take her to be talking about the impending meeting with the labor representative (perhaps informing the assistant that the meeting will be rough), then other features of the context are more important than the clouds outside the window. And then, since people are complex and often have multiple motivations for what seem to be single actions, it would seem that we could interpret the sentence both ways at once.

In general, it is hard to make a case for the claim that a statement about the weather made while the speaker is looking out at the sky is an act of flouting or even violating conversational maxims. We need to turn our attention away from the speaker for a moment and focus instead on the listener. If the listener were to begin the interpretive process by assuming that the only relevant features of context are those pertaining to the meeting, then he or she may come to wonder why the boss is now commenting on the weather (something low on the scale of relevance). In such a case, the listener might take the utterance to fail to live up to conversational maxims, and so would have to find a way to make the utterance relevant. The listener may decide that the best way to do this is to take the remark as really being about the negotiations. (Notice that even with this interpretive structure governed by the assumption that only what pertains to the meeting is relevant, the assistant could stick with the literal and try to find the relevance more directly by asking, "Do you think the storm will make the others late for the meeting?")

---

[29] Grice, "Logic and Conversation," in A. P. Martinich, ed., *The Philosophy of Language* (New York: Oxford, 1985), p. 162–3.

There are two important features of Fogelin's appropriation of Grice. First, as we have seen, twice-true cases show that the same context can support both literal and figurative interpretations. This raises the question whether the indirect speech-act framework, with its supposed required lack of fit with context, is a useful way to distinguish between the literal and the figurative. Second, the Gricean maxims are maxims governing the cooperative exchange of information. Fogelin's contention that the function of both metaphor and simile is to make comparisons puts emphasis on the cognitive and communicative function of figuratively interpreted discourse. This is a natural corrective to old bad classical theories that treat metaphor as a mere rhetorical flourish, but we must not bind ourselves with correctives.

A restraint on audience interpretation that allows the audience to interpret figuratively only if the speaker flouts conversational maxims is too strict. Although we need a view that emphasizes that figurative interpretation may yield information, we do not want to make the conveyance of the information necessary. In the storm-brewing example, the company president may simply have been trying to pass a few minutes before the meeting with simple pleasantries, and only noticed the figurative significance of what she said once she had said it. Or, she may have recognized and enjoyed the fact that the statement could be well interpreted both literally and figuratively. It would be silly to preclude the figurative interpretation simply because the literal interpretation fits. Even if the exchange of information were the most common generic reason for conversation, this surely would have more to do with us and our activities than with what is figurative about the figurative interpretation of discourse. Besides, conversation is just one form of discourse. It has long been a commonplace (though not necessarily a truism) that in literature and poetry the communication of information takes a backseat to other purposes. For all we know, this may be true in daily conversation as well.

In addition to placing less theoretical weight on the communication of information, we need to emphasize that interpretation and the determination of discursive purpose are interdependent. The aptness of a sort of interpretation and the aptness of a particular interpretation are both a function of the purpose of the discourse. But then, the purpose of the discourse is itself a matter of interpretation.

III. CLAIM (0).

(0) Simile and metaphor play the same linguistic role: they make (or prompt us to make) comparisons.

The simile theory of metaphor is misguided. Behind it lies the assumption that similes and metaphors do the same work and that this work is to make comparisons or to get us to make them. Even if we grant this assumption, it provides no reason to believe that the tropes should be given the same analyses. Both my car and my bicycle take me from home to work, but an analysis of how my bicycle manages the task will not explain how my car does it. The argument from similar effect to same cause just is not valid. The simile theorist ignores the role of the description of the effect in the search for its cause. Both an axe and a saw can divide wood, but an axe cannot saw wood and a saw cannot split wood.

The simile theory confuses the task of explaining what a metaphor means with the task of explaining how metaphor (in general) has meaning. It confuses explanation with explication. It may be of heuristic value to treat a particular metaphor (or even every particular metaphor) as a simile in order to fix its meaning but it does not follow from this that metaphor is just implicit simile. And even if metaphor were implicit simile, since the similes constructed from metaphors cannot be understood wholly literally, the simile theory takes us no closer to the comfort of ready-made semantics. There is thus no reason to grant the simile theorist this basic assumption that metaphor is essentially comparative.

Finally, in assimilating simile to literal comparison, the reductive simile theory obscures what is most important to an account of metaphor. The difference between simile and literal comparison lies in the interpretation: '*A* is like *B*' is a simile when interpreted figuratively and a literal comparison when explication or continuation does not involve any figurative interpretation. An adequate account of metaphor will be an account of what it is to interpret a bit of discourse figuratively.

LYNNE TIRRELL

University of North Carolina

REFERENCE AND PRONOMINAL DESCRIPTIONS*

SINCE the appearance of Saul Kripke's seminal article, "Speaker's Reference and Semantic Reference,"[1] there seems to be a growing consensus that the distinction between attributive and referential uses should be regarded as a phenomenon of pragmatics. In this essay, I shall argue that the consensus is mistaken. Moreover, I believe that the mistake is one with considerable significance. The nature of the distinction has been misassessed because Keith Donnellan[2] and others following him have missed an important generalization. That is, definite descriptions have a number of related *pronominal* uses, and referential use is the use of these pronominal descriptions in direct, singular reference. Pronominal uses, I shall argue, contrast systematically and *as a matter of semantics* with instances of attributive use.

One central strand of Kripke's argument against Donnellan runs as follows. Let us assume, as a first hypothesis, that sentences imbedding definite descriptions uniformly have the truth conditions Bertrand Russell proposed. Accepting this hypothesis does not at all preclude us from going on to distinguish two kinds of case. In one type of case ("the complex case"), a speaker utters $\ulcorner$The ϕ is $\psi\urcorner$, believing of some object o he has in mind, that it is [uniquely] a ϕ, and meaning to convey to his audience, by means of his utterance, that o is ψ. In "the simple case," the same sentence is produced, but there is no object about which the speaker has the relevant supplementary beliefs and speech-act intentions. Kripke also stresses that, if sentences containing definite descriptions do have Russellian truth conditions, then it is almost inevitable that both simple and complex cases will arise.

These considerations suggest the following view. First, the *literal meaning* of the speaker's words is always given in accordance with Russell's account, and the special "singular" proposition associated with referential use is treated as part of what *the speaker* meant in uttering his sentence. Second, let us say that *the speaker's referent*

---

* Thanks for very helpful comments to Jeffrey C. King, William W. Taschek, and Mark Wilson.

[1] In *Contemporary Perspectives in the Philosophy of Language*, P. French, T. Uehling, Jr., and H. Wettstein, eds. (Minneapolis: Minnesota UP, 1979), pp. 6–27.

[2] "Reference and Definite Descriptions," *The Philosophical Review*, LXXV (1966): 281–304; see also his "Putting Humpty Dumpty Together Again," *The Philosophical Review*, LXXVII (1968): 203–15.

0022-362X/91/8807/359–87

on a given occasion is the object o (if any) about which the speaker meant to convey that it is ψ. It is denied that the fact that o is the speaker's referent on an occasion renders *o the semantic referent* of the speaker's token of $\ulcorner$the $\phi\urcorner$. It is not the semantic function of the description token to refer to o (if indeed it does) in virtue of o's being the speaker's referent in that context. Within Kripke's framework, the semantic referent of a description is always its unique satisfier. Thus, Kripke, building on Paul Grice,[3] constructs a pragmatic refinement of Russell's theory of definite descriptions, a refined account that purports to offer a better explanation of the data upon which, it has been supposed, a semantical version of the referential-attributive distinction can be grounded. Moreover, Kripke also argues that, if it is assumed that the referential-attributive distinction pertains to the semantics of descriptions, then it is difficult to see how the distinct semantical contributions thus posited are plausibly to be described.[4]

I agree with both of Kripke's conclusions. What I shall attempt to do will be to set out a more complicated range of considerations that do license a semantical distinction between *pronominal* and attributive descriptions. As noted before, Donnellan's referential uses will reappear as a special case of the former. Also, in elaborating these considerations, it will become possible to state a definite proposal concerning the character of the semantical differences between the two general categories.

I

Let us begin with the observation that definite descriptions, like pronouns, may be anaphorically dependent on antecedent terms or phrases. This fact gives rise to a type of ambiguity worthy of attention. The sort of ambiguity I have in mind is easy to illustrate. To begin, consider the statement

(1) Every scientist who was fired from the observatory at Sofia was consoled by someone who knew [him] *the fired scientist* as a youth.

I take it that the italicized description, like the pronoun that could replace it, is a variable bound by the quantifier phrase that fronts the statement. The repeated lexical matter in the descriptor helps here to indicate the tie between the description and the quantifier phrase that binds it. Thus, (1) says

---

[3] For the relevant Grice material, see Part I and essays 15 & 17 in *Studies in the Way of Words* (Cambridge: Harvard, 1989).

[4] Kripke does *not* endorse the pragmatic refinement of Russell's theory briefly sketched above. His main point is that standard considerations in favor of a semantical version of the referential-attributive distinctions are also compatible with the pragmatic alternative, and that the latter seems, in several key respects, methodologically more attractive. He is careful to note in a number of places the choice between these frameworks (and others) is left open by his discussion.

(1) (Every scientist . . . x) (some person y) (y consoled x & y knew x as a youth)

Under certain conditions, however, standard English usage permits the occurrence of a description, similarly bound by the same quantifier phrase, where the lexical indication of the binding is mediated less directly. If a speaker assumes that his audience supposes that

(2) Each scientist who was fired from the observatory at Sofia was a distinguished Bulgarian astronomer.

then, relying on this presupposition, she may utter

(3) Every scientist who was fired from the observatory at Sofia was consoled by someone who knew *the distinguished Bulgarian astronomer* as a youth.

In the imagined context, such an utterance would be little more than a stylistic alternative to (1), i.e., it also states that (1′). What is less clear is how the presupposition (2) of (3) is to be figured into an account of truth conditions. Does (3), under the indicated interpretation, assert the conjunction of (1′) *and* (2), or would the falsity of (2) simply relieve (3) of any truth value whatsoever? Fortunately, present purposes do not require us to resolve this question, but it is an issue that will arise again later.

In any case, the sentence type in (3) clearly has at least one other potential reading. The definite description could be anaphorically independent of the sentence-initial quantifier phrase and occur attributively within the statement. For the time being, let $\ulcorner[\text{the } \phi]_A\urcorner$ mark an attributive occurrence of the description. Then, (3), in such a case, will express

(Every scientist . . . x) (some person y) (y consoled x & y knew [the distinguished Bulgarian astronomer]$_A$ as a youth)

where, I shall assume, the description contributes to the truth conditions of the utterance in the familiar Russellian style.

It should be noted in this connection that the sentence-initial quantifier phrase may be supplanted by a proper name or a demonstrative. Just as we have

(4) Hugo Wexler was consoled by someone who knew *him* as a youth.

we could also have, if the speaker presupposes that Hugo Wexler was a distinguished Bulgarian astronomer, the assertion

(5) Hugo Wexler was consoled by someone who knew *the distinguished Bulgarian astronomer* as a youth.

Modulo the issue about the status of the presupposition of (5), (4) and (5) both say the same thing. Moreover, it seems to me that the pronoun in (4) and the counterpart description in (5) should be treated in a manner parallel to their treatment in (1) and (3)—each pronoun/pronominal description is, in a sense, "bound" by its antecedent. In other words, (1) and (3) are true only if at least one scientist who was fired from the observatory at Sofia satisfies the open sentence

( ) was consoled by someone who knew ( ) as a youth

and, analogously, (4) and (5) are true only if this same open sentence is satisfied by Hugo Wexler. This is to repudiate the idea that, e.g., the pronoun in (4) is a "pronoun of laziness" for the governing name.

Finally, the sentence type (5) exhibits the same ambiguity as the sentence type (3). Clearly, the same description in the same sentence may not "look back" to the name 'Hugo Wexler', and, when it does not, it may have an independent status as an attributive definite description. If so, (5) says

(Someone y) (y consoled Hugo Wexler & y knew [the distinguished Bulgarian astronomer]$_A$ as a youth)

This observation is of interest here since it brings us at least closer to the kind of "referential-attributive" ambiguity that Donnellan has described.

The existence of this kind of ambiguity is undeniable, but it may be felt, with some prima facie justice, that these considerations are so special and so restricted in their application that they represent nothing more than odd exceptions to the more basic rules. But, on the contrary, I believe that these considerations are just the tip of an iceberg: the more fundamental rules are really exemplified here. To substantiate this claim, we need to glance at the intersentential use of pronouns and pronominal descriptions in extended stretches of topic-elaborating discourse.

Some of the key data have become relatively familiar. Here is the beginning of a little story.

(6) Yesterday some man or other left a stick of gum for each of the receptionists. *He* departed unnoticed before *she* discovered *it* on *her* desk.

In a sense that requires explanation, the italicized pronouns in the second statement are introduced by and interpreted in light of the various relevant quantifier phrases in the opening assertion. It has also been noticed that such anaphoric pronouns can often be replaced by suitable descriptions.[5] Thus, in (6), we could have had 'the man', 'the stick of gum', and 'the receptionist' instead of the pronouns that actually occur. Indeed, in some instances, the use of descriptions rather than pronouns becomes obligatory if the intended anaphoric ties are to remain clearly delineated. If (6) had started, 'Some man or other left a stick of gum for some other man at the office', then reduction of unnecessary ambiguity would call for a continuation of the discourse such as

> *The man who left the stick of gum* departed unnoticed before *the man with whom it was left* discovered it.

It is important to emphasize, as (6) implicitly indicates, that these pronouns and pronominal descriptions can "look back" to universal quantifier phrases. Consider

(7) Any novel Bruno wrote was read by each female critic on the staff.

where this discourse is extended by

(8) She hated it with a passion.

With (7) as its "topic sentence," (8) says

> All of Bruno's novels were hated with a passion by every female critic on the staff.

Similarly, if (7) had been

(7') Any novel Bruno ever wrote was read by at least one female critic on the staff.

then (8), in that context, would have expressed

> Each of Bruno's novels were hated with a passion by at least one female critic on the staff (who read it).

As in our previous example, one or both of the pronouns in (8) might just as well have been a pronominal description instead. A satisfactory account of these terms should provide a unified treatment of the pronouns *and* pronominal descriptions in their relations *both* to existential and to universal quantification.

---

[5] The point is noticed by Gareth Evans in "Pronouns, Quantifiers, and Relative Clauses (I)," in *Collected Papers* (New York: Oxford, 1985), pp. 76–152; and I strongly emphasize the issue in my "Pronouns and Pronominal Descriptions: A New Semantical Category," *Philosophical Studies*, XLV (1984): 1–30.

Life and our present exposition would be simple if we could suppose that these pronouns/pronominal descriptions are also variables bound by the quantifier phrases that introduce them into discourse. I take it, however, that it is widely and correctly agreed that this simplest hypothesis is wrong.[6] Unfortunately, the question of a proper semantics for such terms is a topic of considerable controversy in contemporary debate. In what follows, I shall sketch a view that Jeffrey King and I[7] have developed more extensively elsewhere.

It is striking that the pronouns and pronominal descriptions under scrutiny mimic much of the behavior, syntactic and semantic, of the parameters or "arbitrary names" that are standardly deployed in systems of natural deduction.[8] In Benson Mates's[9] system, for example, the following trivial deviations may be constructed:

D_1				D_2			
$\langle 1\rangle$	(1)	$(\exists x)Fx$	P	$\langle 1\rangle$	(1)	$(x)Fx$	P
$\langle 2\rangle$	(2)	$(x)(Fx \rightarrow Gx)$	P	$\langle 2\rangle$	(2)	$(x)(Fx \rightarrow Gx)$	P
$\langle 3\rangle$	(3)	Fa	P	$\langle 1\rangle$	(3)	Fa	UI
$\langle 2\rangle$	(4)	$Fa \rightarrow Ga$	UI	$\langle 1, 2\rangle$	(4)	$Fa \rightarrow Ga$	UI
$\langle 2, 3\rangle$	(5)	Ga	MP	$\langle 1, 2\rangle$	(5)	Ga	MP
$\langle 2, 3\rangle$	(6)	$(\exists x)\,Gx$	EG	$\langle 1, 2\rangle$	(6)	$(x)Gx$	UG
$\langle 1, 2\rangle$	(7)	$(\exists x)\,Gx$	EI				

The term '*a*' in D_1 and D_2, introduced into both derivations as a kind of *schematic singular term,* is an instance of such a parameter. Although it has not been the usual practice to do so, it is possible to assign intuitively correct truth conditions to an arbitrary line containing one or more of these parameters, and, for a well-conceived system of natural deduction, this can be carried out in such a way that each correctly derived formula with parameters is a logical consequence of the premises upon which it rests. The salient fact for present purposes is that the truth conditions of a line with parameters is determined by the (schematic) meaning of the formula *and* by the specific context of the line within the structure of the derivation. Derivational context helps to fix the truth conditions of a line 1 by showing the way in which the parameter in 1 has been introduced (i.e., by EI or UI), by manifesting the relations within the derivation of distinct parameters that occur in 1, and by exhibiting the premises and temporary suppositions from which 1 has been derived. We will notice, if we examine line 5 of D_1 and D_2, respectively, that the formula ⌜*Ga*⌝ (with the interpretation of ⌜*G*⌝ held fixed) in D_1 has the force of

---

[6] For arguments against the "bound-variable" view, see the papers cited in fn. 5, and Jaakko Hintikka's "Quantifiers in Natural Languages: Some Logical Problems," in *Game Theoretical Semantics,* Esa Saarinen, ed. (Dordrecht: Reidel, 1978). As noted in the text, I am under the impression that many agree that such uses of pronouns in discourse are not to be treated as bound variables.

[7] "Pronouns, Descriptions, and the Semantics of Discourse," *Philosophical Studies,* LI (1987): 341–63; and my "Pronouns and Pronominal Descriptions."

[8] For an alternative view of the semantics of parameters, see Kit Fine, *Reasoning with Arbitrary Objects* (New York: Blackwell, 1985). Although I am inclined to reject Fine's apparatus of "arbitrary objects," this book is filled with rich and suggestive observations and analyses of the pertinent material. For a comparison of the approach King and I favor to Fine's, cf. King, "Anaphora, Instantial Terms, and Pronouns," *Philosophical Studies,* LXI (1991): 239–65.

[9] *Elementary Logic* (New York: Oxford, 1972, 2nd ed.).

In each instance (for ⌜*a*⌝), ⌜*Ga*⌝ is true.

while the same formula in D_2 conveys

In at least one *F* instance (for ⌜*a*⌝), ⌜*Ga*⌝ is true.

In other words, a formula like ⌜*Ga*⌝ in these contexts exhibits a systematic ambiguity analogous to the ambiguity of (8) above when it arises out of either (7) or its variant (7′). Much more complicated examples can easily be constructed which involve either parameters in a derivation or pronouns/pronominal descriptions in a discourse, but our simple examples should suffice for the issues before us now.

It is my view, of course, that these pronouns and pronominal descriptions also operate as parameters or schematic singular terms within the confines of a topic-structured conversation, exposition, or narrative. In interpreting such terms, we look to the manner in which they have been introduced as dependent upon antecedent claims and assumptions. When we speak of "a discourse" here, we should include not only the actual utterances, connected by relevance to a larger subject, but also those mutually shared assumptions which the speakers have taken for granted as the conversational framework that guides the series of responsive remarks that they produce. Where we have a pronoun or pronominal description serving as a parameter, we have to trace it back to its spoken or merely tacit source. Often, as in our earlier examples, that source will be a quantifier phrase in some earlier statement or presupposition of the discussion.

These natural-language parameters are variable-like terms that serve in their sentential settings to express appropriate *general* propositions. But they differ from the standard, bound variable of formal logic in at least three important respects. First, they are not literally bound by an antecedent quantifier phrase, but in virtue of context they are, so to speak, semantically unfree. Second, they are descriptionally enriched. If a description ⌜the ϕ⌝ is used as a parameter, the predicables in its description constrain the kind of interpretation the description can receive. In virtue of the fact that English third-person singular pronouns are marked for gender, the same observation holds in a more modest way for them. (I shall consider one significant aspect of this property shortly.) Third, as adumbrated above, extrasentential discourse context determines various features of the proper interpretation of a parameter in a given utterance. That is, (i) the discourse will associate with each parameter a *mode of generality*, either universal or existential as the case may be. Thus, if (8) is uttered as an extension of (7′), then 'she' in (8) is associated with existential generality, while 'it' in (8) is universally general. In the same way, the parameter '*a*' in the derivations is associated with different modes of generality depending upon whether it has been introduced by UI or EI. (ii) If a statement contains more than one parameter, then the discourse will determine a scope-like *ordering of logical precedence* among them. Looking at (8) again in relation to (7′), the knowledge that 'she' is associated with existential generality and 'it' with universal generality does not tell us whether (8) says that each of Bruno's novels was hated by some of the female critics or that some female critic hated all of his novels. So context must and generally does settle this matter as well. Finally, (iii) discourse determines the *domains* over which the pertinent parameters range. If (7) leads into (8), the domain of 'she' in (8) is constituted by the female critics on the staff. The domain of 'he' in the second statement of (7) is made up of the men who left a stick of gum for each receptionist.

In even moderately more complex cases, however, we have to consider the domain of a parameter in a statement *with respect to* an assignment of individual values to other parameters upon which the target parameter depends.[10] Suppose a discourse begins with

(9) (Some x) (some y) Fxy

and the parameters ⌜a⌝ and ⌜b⌝ are introduced as instantiations of the variables 'x' and 'y' in (9). Having done so, the speaker goes on to assert, e.g., ⌜Gab⌝. Intuitively, the domain of ⌜a⌝ is just the set of items that bear F to at least one thing. But "the" domain of ⌜b⌝ should be the set of things to which a bears F. And yet, this condition makes no sense as long as ⌜*a*⌝ remains a purely schematic term. We obtain a definite domain for ⌜*b*⌝ only after some object from ⌜*a*⌝'s domain is assigned to ⌜*a*⌝ as its referent. In other words, "the" domain of ⌜*b*⌝ *depends upon* a choice of a value for ⌜*a*⌝, and the concept that we need is: the domain of ⌜*b*⌝ with respect to an assignment to ⌜*a*⌝. This allows us to specify correctly that ⌜*Gab*⌝ is true in the context if

---

[10] For a much more extended discussion of this crucial concept of "(parameter) *dependence*," see the book by Fine and the article by King mentioned in fn. 8. My manner of introducing this concept, briefly sketched below, is somewhat different from theirs.

> (Some x) ($x \in$ the domain of ⌜*a*⌝) & (some y) ($y \in$ the domain of ⌜*b*⌝ with respect to x & Gxy)

Therefore, when (7′) provides the context for (8), the schema for a domain for 'she' in (8) is ⌜female critic on the staff who read ( )⌝. A particular domain is defined when one of Bruno's novels is taken as the value for ⌜( )⌝.[11]

When all of these features—mode of generality, ordering of logical precedence, and independent and dependent domains—have been adequately aligned by the discourse, then a parameterized statement assumes definite and definable truth conditions.[12] I should mention that there is nothing unprecedented about the basic mechanisms at work in this affair. The sentence, 'Elephants fly' is, as it stands, ambiguous or indeterminate between (at least) 'All elephants fly' and 'Some elephants fly'. On a given occasion, context, including the speaker's relevant intentions, will decide the wanted mode of generality. Similarly, the sentence 'Every man loves some woman' is also notoriously ambiguous, and contextual factors must indicate the order of logical precedence between the pair of quantifier phrases. Finally, when I say, "Every napkin is frayed," it will be context that further restricts the range of quantification to be, e.g., napkins in the house, napkins on this table, or whatever. In the case of the parameters, it is simply that all of these features are fixed simultaneously, and the contextual factors that effect the triple determination tend to be structurally more complex. Both the nature and the structural complexity of these discourse contributions are more perspicuously exhibited in the regimented environment of natural deduction systems.

---

[11] These brief remarks are meant only to give an impression of the special semantical complexity that attaches to parameters, but to do so in a way that keeps in focus their relation to the simple and more familiar machinery of quantifiers and the variables they bind.

[12] For a developed semantics for parameters in a system of natural deduction elaborated along the lines I favor here, see King's "Pronouns, Descriptions, and the Semantics of Discourse," and his "Anaphora, Instantial Terms, and Pronouns."

These remarks on pronouns/pronominal descriptions qua parameters are too brief—the topic is a rich and subtle one. They should be enough, however, to extend the central line of thought I was developing earlier. For even if pronominal descriptions are used sparsely as bound variables, their use as parameters that express generality is certainly widespread. Moreover, this use generates new instances of the sort of ambiguity that was illustrated when bound pronominal descriptions were on the stage. Modifying a previous example, we have

(10) (a) At least one } scientist at the observatory at Sofia
(b) Every } was fired.

where either of these is continued by

The fired scientist was consoled by someone who knew him as a youth.

But now, let us imagine once more that the speaker presupposes (2), i.e.,

Every scientist who was fired from the observatory at Sofia was a distinguished Bulgarian astronomer.

If so, then the elaboration of (10a) or (10b) could take the form

(11) The distinguished Bulgarian astronomer was consoled by someone who knew him as a youth.

With the pronominal description read as a parameter that is governed by the presupposition (2), (11) says, in relation to (10a),

(At least one scientist . . . *x*) (some person *y*) (*y* consoled *x* & *y* knew *x* as a youth)

or in relation to (10b),

(Every scientist . . . *x*) (some person *y*) (*y* consoled *x* & *y* knew *x* as a youth)

In contrast with both of these readings, however, there is another in which the pronominal description is anaphorically independent of (10) and occurs attributively in (11). On this reading, (11) states,

(Some person *y*) (*y* consoled [the distinguished Bulgarian astronomer]$_A$ & *y* knew him as a youth)

This type of ambiguity is, it seems to me, quite commonly exemplified.

Given this parallel between pronominal descriptions and third-person pronouns, and given that the pronouns have a ubiquitous *deictic* or *indexical* role, one would expect that pronominal descriptions might also share this role. And, of course, it seems as if they do. When a speaker gestures toward Hugo Wexler in conversation with another person and says,

> The distinguished Bulgarian astronomer is being consoled by someone who knew him as a youth.

then we have an apparent instance of a pronominal description used deictically.

It is precisely the semantic status of such utterances, however, that has been the subject of dispute. Let

(12) The $\phi_{\downarrow}$ is ψ

represent a deictic employment of the description, and let o be the object that is contextually indicated as the subject to which the predicate is purportedly ascribed. The parallel with pronouns suggests that the deictic character of the description *is semantically relevant* to the strict and literal meaning of the utterance that contains it. That is, the semantic relevance hypothesis (SRH) claims that the strict and literal meaning of (12) on the relevant occasion is such that (12), in that context, expresses the Russellian singular proposition

of o, that it is ψ

while depending upon the singular presupposition

of o, that it is also ϕ

There is an issue here, as there has been in previous cases, about how the presupposition is to be counted in relation to the strict and literal meaning of the utterance, and in particular, about how the presupposition acts upon the conditions of truth and falsity. For myself, it seems that if o is not ϕ, then what is strictly and literally said by the words should be assessed as untrue, but I am not certain whether the untruth is simply false or is defective in relation to truth in some other way. Nevertheless, however this question may be decided, the world-to-world truth conditions of (12) will differ, according to (SRH), from those of an utterance of ⌜The ϕ is ψ⌝ in which the description is attributive. Obviously, deictic uses of pronominal descriptions correspond closely to the occurrences that Donnellan called "referential uses of definite descriptions."

(SRH) seems immensely natural from within the larger framework I have sketched. In fact, it is plausible to maintain that deictic uses of pronouns and pronominal descriptions constitute a limiting case, albeit a central limiting case, of their alternative uses as variables and parameters. Suppose that discourse context can assign to a parameter in a statement—not a mode of generality and domain—but a particular object, an object that is thereby assigned to the parameter as the subject of the conjoined predicate. Indeed, this is exactly what seems to take place if we consider an analogue of the discourse (10) + (11) in which the initial quantifier phrase is replaced by a proper name, e.g.,

> Hugo Wexler was fired from the observatory. The distinguished Bulgarian astronomer [or: He] was consoled by someone who knew him as a youth.

I take it that the second statement, in virtue of the intended connection between the name and the pronominal, says *of* Hugo Wexler that he was consoled . . . , etc. Therefore, it is difficult to see why sundry nonlinguistic features of the speech context should not be able to achieve the same result. A demonstrative gesture accompanied by a plain intention to be speaking of the entity demonstrated could yield a similar assignment to the pronominal parameter and so could other less explicit, more elaborated facets of the speaker's discourse circumstances.

In many systems of natural deduction, the same set of terms is used, with a studied and harmless ambiguity, both as parameters and as individual constants. In any particular derivation, such a term will function in one way or the other, but, from derivation to derivation, the function of a given term may well change. Why should it not be the case that English, with its pronouns and descriptions, incorporates terms that behave similarly? Why should English not have expressions that appear as schematic singular terms in one discourse and then pop up in others as genuine singular terms referring to specific objects? Why not, indeed? I believe that this is precisely the case.

We still have not settled on the truth of (SRH), however. Despite its appeal, the fact remains that Kripke has produced powerful arguments to undermine its various attractions. Therefore, in the next two sections we turn to an investigation of those arguments and of some significant related questions that they raise.

II

In "Speaker's Reference and Semantic Reference," Kripke several times underscores the methodological principle that a "unitary" semantical account of definite descriptions is preferable to any theory that postulates an intuitively dubious semantical ambiguity. The principle is surely correct, and Kripke is right to emphasize its importance. But, from our present point of view, we see that the principle points inconclusively in different directions. We have reminded ourselves that no unitary account in Russell's mode will accommodate all of the interesting, systematic uses of descriptions: the uses of descriptions as bound variables and parameters cannot be so accommodated. Any overview of the data must recognize some broad distinction between descriptions used attributively and descriptions used pronominally.

A defender of Russell will contend, however, that the pragmatic refinement of Russell's theory offers a unitary account of all nonanaphoric uses of description. And, it will be added, this is all that the theory could reasonably be expected to cover. On the other hand, (SRH) contests even this more restricted claim. Deictic uses of pronominal descriptions should be recognized as distinct from the attributive uses, and the deictic use forms a part of a unitary theory of pronominal descriptions. Or, putting the matter in a different way, according to (SRH) the deictic uses are themselves, in a sense, anaphoric. Unlike the standard examples of anaphora, they are anaphorically dependent upon various nonlinguistic contextual determiners of singular reference which are present in or available from the utterer's speech-act situation. Hence, both theories propose a kind of unification of the pertinent phenomena. The difficult question is: "Which form of unification is correct?"

The natural place to look for evidence to decide this question is the plentiful domain of more or less obviously *incomplete* descriptions. Someone says, "The baby is crying" or "The table is covered with books," and it is not a part of what has thereby been said that there is one and only one baby or one and only one table. On the face of the matter, many such statements *seem* to involve deictic, pronominal descriptions, and this looks favorable to (SRH). Nevertheless, it is the standard reply of the Russellian that Russell's analysis is not always to be applied naively to the bare words actually emitted by the speaker. Rather, it is supposed that the speech context will somehow supply a restrictive condition upon the truncated description, a restriction that makes the implication of uniqueness something the speaker could plausibly have meant.[13] Thus, both

---

[13] I should note that Kripke himself sets the issue of incomplete descriptions to one side in his essay, and he does not endorse the emendation of Russell's ideas I shall be discussing. He states: "Contrary to the Russell picture, I doubt that such [incomplete] descriptions can always be regarded as elliptical with some uniquely specifying conditions added. And it may be the case that a true picture will resemble Donnellan's in important respects" (*op. cit.*, pp. 6–7). In what follows, I shall argue that Kripke's doubts are well-founded.

(SRH) and Russell's account, emended to cover patently incomplete descriptions, invoke a regular role for context in determining what is literally said by "referential" tokens of $\ulcorner$The ϕ is $\psi\urcorner$, but the character of what is so determined differs substantially between the two proposals.[14]

Up to a point, contextual qualification of some incomplete descriptions can and should be countenanced. For instance, speakers often leave various obvious indexical specifications unstated but easy to supply. When Donnellan's party guest utters "the man who is drinking champagne," it will be clear that he means "the man who is *now* drinking champagne *here*" or something of the sort. In particular, there are many occasions on which a description used attributively will be patently incomplete. A detective, contemplating the horribly mutilated corpse, says "The murderer is insane," but context makes it plain that he is to be understood as asserting "The murderer *of Smith* [or: The murderer of him] is insane."[15] These commonplace styles of supplementation, however, will not generally suffice to yield a uniquely satisfied descriptor. Speaking of a certain piece of furniture in an antique store full of tables, someone states, "The table is covered with books." Here also the Russellian requires that some suitable emendation be recoverable from context, and, at this juncture, it seems to me, the idea is protected only by a convenient fogginess about how context is to do its job.

Presumably, the usual thought has been that there will be a particular descriptive restriction that the speaker has in mind and that he intends to qualify his audible definite description. Now, this vague thought seems to me to be almost certainly wrong, but it is one that admits of so many possible variations that I shall simply attempt to explain why the prospects of any of its versions look quite bleak. The upshot will be that this doctrine is exposed to many of the problems that afflict the similar view that the semantic referent of a proper name token is determined by an associated description.

---

[14] I do not have space to discuss views, like Kent Bach's, that radically reject the need for *any* contextual supplementation of the literal meaning or "content" of utterances that contain incomplete definite descriptions. See his *Thought and Reference* (New York: Oxford, 1988), pp. 91–130. On such a view, what has been literally expressed by such an utterance will be, in general, simply false. On the other hand, what the speaker thereby means to convey may well be true. For a critical discussion of Bach on this issue, see William W. Taschek's review of the book in this JOURNAL, LXXXVII, 1 (January 1990): 38–45. I shall note only that it seems highly implausible that, when the proud father says to his wife, "The baby is prettier than any other babies I've ever seen," his words, on that occasion, *literally express a contradiction.*

[15] Scott Soames emphasizes the importance of these considerations in "Incomplete Definite Descriptions," *Notre Dame Journal of Formal Logic,* XXVII (1986): 349–75.

First, there are cases in which a token description intuitively refers to a particular object even though the description is not uniquely satisfied and the speaker has *no resources at all* for qualifying his words adequately. Admittedly such cases tend to have a somewhat *recherché* character, but they nonetheless exist. Suppose that Jean Valjean, as a very old man, has only the spottiest memory of his life up to a year before the events now to be described. One day while walking through Paris he catches sight of a face that fills him with fear and foreboding. When he sees the face he cannot say what it is that he so dreads, and he continues through the day brooding over his troubled state of mind. By later that evening two things have developed. He has totally forgotten that it was the sighting of the face that gave rise to his continued emotional disturbance and, at the same time, he comes to be seized with the conviction that, as he expresses it, "My old enemy is back in Paris." In fact, it is this thought that he now takes to give the object of his fear.

Although Valjean insists that there is a definite person of whom he is thinking, he admits that he has had a number of enemies in the past and he cannot produce any further description to single out the person in question. Finally, imagine that we know, although Valjean does not, facts such as these. The original cause of Valjean's frightened response was the fact that the face he saw bore a definite resemblance to the face of his old enemy Javert. Further, only an intimation, conscious or unconscious, of Javert's presence could generate in Valjean this specific syndrome of intense emotional disturbance. Finally, if Javert, as Valjean had known him years ago, could be directly presented to Valjean now, then Valjean would realize straight off that *this person* is the old enemy he means. In other words, despite the fact that Valjean is unable to summon to mind any helpful representation of Javert, he would be able to recognize the Javert of his past as the man he has been thinking of all along. As facts of this kind are assembled, it becomes hard to deny that Valjean's use of 'my old enemy' does refer to Javert even though nothing like a uniquely identifying description has been provided.

Perhaps it may seem that Valjean has access at least to the description 'the old enemy of mine who is the cause of my present fear'. Let us add to our story, however, one of those amazing coincidences typical of a 19th century novel like *Les Misérables*. The man whose face reminds Valjean unconsciously of Javert is also a former enemy of Valjean's, though Valjean does not register that further fact in any way. Thus, the emended description does not identify Javert uniquely. No doubt there are various other emendations that could be tried, but I think that this last attempt is typical. If the expanded description is not question-begging, a simple alteration of the story will render it improper.

Second, especially in situations where the object referred to is well-known to the speaker and to his audience, and this fact is mutually known, there will be a wide range of qualifications which the speaker is in a position to offer but no one of which is something the speaker clearly had in mind as meant to qualify his words. Also, some of these potential qualifications may yield a uniquely denoting description but many others will not, or, at least, the qualified descriptions will not uniquely denote the intuitive referent. The Russellian must therefore find considerations that *determine* which of the available qualifications is to play the role, as he sees it, of determining the relevant designation. To this challenge there is, I suspect, no satisfactory solution.

If one takes the qualification the speaker would have given at the time of utterance if he had been asked $\ulcorner$Which ϕ?$\urcorner$, then—assuming (wrongly) that the counterfactual here is, in general, well-defined—there will be examples in which the emended description, but not the original, misdescribes the ϕ in question. A man says to his wife, "The girl ought to be punished," and he would naturally and spontaneously fill this in as 'the girl who is our daughter'. But, of course, he can be mistaken about his relationship to the girl. Even if he is allowed to weaken his favored elucidation to 'the girl whom I believe to be our daughter', this description will be improper if the wife has managed a perverse domestic juggling of identical twin girls. There are many cases in which it is clear which object is referred to and what facts determine that object to be the referent, but where it is totally obscure what further identificatory phrase is supposed to qualify the uttered description in the way the Russellian demands. The most obvious choices will sometimes yield intuitively wrong results, and it is difficult to surmise what plausible but less obvious criterion of "qualification determination" could be advanced.[16]

[16] I also shall not explicitly discuss alternative Russellian approaches that do not treat incomplete descriptions as essentially elliptical but do require context to provide a semantically relevant contextual supplementation. It has been held that speech context can supply a *domain* of individuals D such that the relevant utterance is true just in case there is one and only one object that is ϕ in D and it is ψ. As a rule, D is not determined through an implicit supplementation of the patently incomplete description. Versions of this idea may be found in Jon Barwise and John Perry, *Situations and Attitudes* (Cambridge: MIT, 1983); and Jaakko Hintikka and Jack Kulas, *Anaphora and Definite Descriptions* (Dordrecht: Rei-

The package that consists of the pragmatic extension of Russell's theory combined with a Russellian treatment of incomplete descriptions faces a related, serious difficulty. Each of the sentences

(12a) He who murdered Smith is insane.
(12b) This man who murdered Smith is insane.
(12c) That man who murdered Smith is insane.

contains subject terms that have attributive or attributive-like uses. The sentences, in appropriate circumstances, may be used so that they are strictly equivalent to—in fact, so that they are stylistic variants of—the sentence

(14) The man who murdered Smith is insane.

where this description occurs attributively. Such a use of (12b) and (12c) will perhaps be most natural in a situation when the discussion has reverted to, for example, previous speculations about who may have murdered Smith. Now, *if* the Russellian proposals concerning incomplete descriptions were to succeed, then it is hard to see why the same proposals should not work for demonstratives as well. Let

(14a) He$_{\downarrow}$ (she$_{\downarrow}$, it$_{\downarrow}$) is ψ
(14b) This $\phi_{\downarrow}$ is ψ
(14c) That $\phi_{\downarrow}$ is ψ

be statements containing demonstratives that have been uttered in a context *C*, and let *o* be the item referred to in *C* by means of the relevant demonstrative. One might then suppose that *o* is the semantic referent in *C* of the demonstrative and that the imbedding utterance strictly and literally expresses the proposition

(15) of *o*, that it is ψ

A new application of the "package" under scrutiny will inform us, however, that this is wrong. For these proposals will tell us that the subject terms in deictic tokens of (14) are really incomplete versions of the expressions 'He who . . .', ⌜This ϕ who . . .⌝, and ⌜That ϕ who . . .⌝, where a suitable completion of the material elided will be indicated by the speech context. If ⌜θ'd⌝ represents that completion, then the strict and literal meanings of (14) will be given by

---

del, 1987). I believe, however, that many of the issues I raise in the text against the "supplementation" view are easily extended to generate problems for these alternatives. For an incisive critical discussion of the specifics of the Barwise-Perry treatment of incomplete descriptions, see Soames, "Incomplete Definite Descriptions."

He who θ'd is ψ
This ϕ who θ'd is ψ, etc.

with the subject terms interpreted attributively in each case. On this view, the Russellian singular proposition (15) is to be regarded as a content that the speaker, through his utterance, merely implicates.

I assume that there will be wide agreement that this is an overzealous application of the pragmatic machinery Kripke invokes. Ignoring the difficult case of proper names, this application of the machinery will leave, at best, only 'I' and some purely indexical uses of terms like 'here' and 'now' within the category of terms that refer directly to their semantic referents. Although Russell himself might have been happy with such a conclusion, the result is, I take it, quite unsatisfactory. Therefore, the unrestrained proponent of the Russellian package is faced with a dilemma. Either he gives up Kripke's pragmatic revision of Russell from the beginning, or he rejects the Russellian doctoring of incomplete descriptions. If the second option is adopted then he has no account, alternative to (SRH), of "deictic" incomplete descriptions. And, finally, since the incompleteness of a description cannot by itself be crucial to its *semantic* status, he had better acquiesce in a semantic account of deictic uses of pronominal descriptions that are, as it happens, satisfied uniquely.

Thus far, all we have assumed about attributive descriptions is that statements containing them have Russellian truth conditions. But Russell's own investigations of definite and indefinite descriptions suggest a further thesis: the thesis that ⌜the ϕ⌝, when it occurs attributively, is a quantifier phrase of English.[17] If ⌜The ϕ_A is ψ⌝ has the truth conditions of ⌜Something or other which is uniquely ϕ is ψ⌝, then it seems a natural and modest hypothesis that ⌜the ϕ_A⌝ just *is* a quantifier phrase that means (roughly) ⌜something or other which is uniquely ϕ⌝. Therefore, ⌜the ϕ⌝, synonymous with ⌜the *only* ϕ⌝, is to be classified with (nonanaphoric and nondeictic uses of) the following: ⌜the ϕs⌝, ⌜the few ϕs⌝, ⌜the several ϕs⌝, ⌜the n ϕs⌝, and ⌜the many ϕs⌝. These all are English quantifier phrases, the first in the singular, the rest in the plural. Let us adopt the convention of writing $\ulcorner . . . Q\phi . . . \urcorner$ as $\ulcorner (Qx: \phi)(. . . x . . .)\urcorner$. Now, our present proposal, combined with our previous discussion, tells us, in effect, that, if $\ulcorner$The ϕ is $\psi\urcorner$ contains an attributive description, it is properly represented by

---

[17] I shall not attempt here any positive defense of the view that descriptions, used attributively, are quantifier phrases. Among recent writers on the subject, the view is endorsed both in Evans, and Stephen Neale, *Descriptions* (Cambridge: MIT, 1990). For Evans, the descriptions are *binary* quantifiers; for Neale, they are *restricted* quantifiers. For present purposes, *any* view of the semantics of definite descriptions will do which (i) yields the result that they engender Russellian truth-conditions and (ii) do not belong in the category of definite singular terms. For more on the issues raised by this second condition, see the conclusion of the following section.

(The x: ϕ) ψx

while, if the contained description is pronominal and deictic, the statement has the form of

ψ(the $\phi_\downarrow$)

where $\ulcorner$the $\phi_\downarrow\urcorner$, as before, is an indexical singular term. For the remainder of this paper, I shall take the phrase 'the semantical version of the referential-attributive distinction' to refer to this hypothesis.

This or any similar proposal remains suspect, however, in the face of intuitions that the meaning of the definite article is univocal in both of the two uses. Given the proposal above, $\ulcorner$the $\phi\urcorner$, when it serves as a quantifier phrase, carries the uniqueness condition, but when it is a pronominal description used deictically it does not. Conversely, a deictic pronominal description purports to refer rigidly to a particular item, but the quantifier phrase does not. In the light of contrasts like these, one would expect there to be a reasonably sharp linguistic intuition that the meaning of the words, e.g., "the murderer of Smith" is different between the two uses. But no such *definite* intuition seems to be forthcoming. Are we not in danger of postulating a semantic ambiguity that our basic semantic competence in the use of descriptions does not plausibly sustain?

I think that these remarks express at least some of the concerns behind a methodological consideration that Kripke brings to bear against a semantic account of the referential and the attributive. He states,

> . . . 'Bank' is ambiguous; we would expect the ambiguity to be disambiguated by separate and unrelated words in some languages. Why should the two separate senses be reproduced in languages unrelated to English? First, then, we can consult our linguistic intuitions, independently or any empirical investigation. Would we be surprised to find languages that used two separate words for the two alleged senses of a given word? If so, then, to that extent our linguistic intuitions are really intuitions of a unitary concept, rather than of a word that expresses two distinct and unrelated senses (*op. cit.*, p. 19).

A little later, he continues,

> . . . I at least would find it quite surprising to learn that, say, the Eskimo, used two separate words "the" and "ze", for the attributive and referential uses. To the extent that I have this intuition, to that extent I think of "the" as a unitary concept (*op. cit.*, p. 19).

How can we explain the apparent upshot of this intuitive test?

I think that our continuing analogy between pronouns and pronominal descriptions is quite instructive in relation to these last questions. Consider the following examples:

> If any man marries too young, *he* will regret it.
> Some cowboy stopped at the camp. *He* rode a horse.
> Every cowboy worked the range. *He* rode a horse.
> He_1 rode a horse.
> *He* who murdered Smith is insane.

and, exploiting a common idiom,

> George Eliot was not a *he*.

Suppose one were to ask whether the word 'he' in these statements does or does not have the same lexical meaning. It is my impression that the query is likely to lead to confusion and, perhaps to an answer like, 'Well, it does and it doesn't'. The problem arises, it seems to me, because our intuitive concept of "lexical meaning" is too crude to do well in such a case, and it is this crudity which confuses the linguistic intuitions we try to describe using the defective concept.

Up to a point at least, it is easy to make out where the difficulty lies. On the one hand, the semantic function of the pronoun is different in all or most of these examples. In that respect, the "meaning" of the pronoun can be said to vary from one statement to the other. On the other hand, the descriptive conditions or constraints upon the correct "application" of the pronoun remain constant through all of these occurrences. That is, the term "applies" correctly to a potential subject of application only if that entity is a male. No doubt, the word 'applies' in this characterization is either ambiguous or it is an umbrella term that covers all of the pertinent semantic roles. Still, how it is to be understood in each case is clear enough. Therefore, in this respect, we might well be led to say that 'he' has the same lexical meaning in each of our last examples. (We would not expect to find a separate dictionary entry for all or even most of these uses.)

In considering these statements, we see why a standard example like 'bank' gives us a bad model for grasping the "ambiguities" before us now. 'Bank', used in relation to financial institutions and sides of rivers, retains the same kind of semantical role. In both of its senses, the word 'bank' is a predicable—more specifically, a general term. The two senses arise because the term is conventionally associated with two sets of applicability conditions. This is the reverse of the situation with regard to the pronoun. The word 'he' carries its minimal conditions of applicability along as it systematically shifts the semantical function it performs in the utterances in which it appears.

It is my view that we ought to look at definite descriptions in *their* various semantical roles, in the same way. A given description does not change the conditions of correct application which it semantically expresses in the descriptor even when it does change from one pronominal function to another or from its pronominal status to its attributive use as a quantifier phrase. The "descriptive sense" (as we might put it) of, e.g., 'Smith's murderer' is just the same no matter whether the term occurs as a quantifier or as a pronominal description of one stripe or another. In this way, we can accept and explain the intuition that the descriptive phrase expresses "a unitary concept" while maintaining, at the same time, that very different semantical roles should be distinguished.

I have not attempted a direct answer to Kripke's question about whether we would expect pronominal and attributive descriptions to translate into distinct linguistic forms in languages other than English, and I have not done so for a reason that he himself supplies. In a footnote, he briefly considers some ways in which his intuitive test "needs further explanation and refinement." He says,

> . . . a word may have different senses that are *obviously related* [the italics are mine]. One sense may be metaphorical for another (though in that case, it may not be a separate sense, but simply a common metaphor.) "Statistics" may mean both statistical data and the science of evaluating such data and the like. The more we can explain relations among senses, and the more 'natural' and 'inevitable' the relationship, the more we will expect the different senses to be preserved in a wide variety of languages (*op. cit.*, p. 26, fn. 29).

I have tried to indicate why I believe that gendered pronouns and definite descriptions in English provide beautiful, distinctive, and fairly complex examples of the situation Kripke describes in this passage. The terms in question have a family of semantical functions that are closely and naturally related to one another, and, as a consequence, our expectations about how other languages will distribute these functions within their lexicons are, on the whole, pretty dim. We should not allow the nagging feeling that *somehow* the "lexical meaning" of descriptions is "the same" dissuade us from the conclusions we have drawn.

III

In this section, we turn first to an important and influential objection that Kripke raises against any semantical competitor to the pragmatic refinement of Russell's theory. A fairly careful examination of this objection will help considerably to clarify some of the major points that divide the semantical version of the referential-attributive distinction from its quasi-Gricean alternative.

Kripke begins by pointing out that there are simple and complex cases involving *proper names*. An instance of a complex case is described in the following passage:

> Two people see Smith in the distance and mistake him for Jones. They have a brief colloquy: 'What is Jones doing?', 'Raking leaves.' 'Jones' in the common language of both is a name of Jones, it *never* names Smith. Yet, in some sense, on this occasion, clearly both participants in the dialogue have referred to Smith, and the second participant has said something true about the man referred to if and only if Smith is raking leaves. . . . How can we account for this? (*op. cit.*, p. 14).

In the corresponding simple case, there is no man seen in the distance. The speaker says, simply using 'Jones' as a name of Jones, 'Jones is raking leaves'. This utterance might express, for instance, an educated guess about what Jones, wherever he is, is doing at that time. Considering the two cases, Kripke points out that it lacks all appeal to account for the differences between the cases by postulating a semantic ambiguity either in the name or in the sentence that contains it. In both situations, the semantic reference of the name is Jones and the sentential utterances say, strictly and literally, *of him* that he is raking leaves. In the complex case, because of the speaker's conflation of Smith with Jones, it is also true that the speaker intended *of Smith* that the name token refer to him. But, although *the speaker* was, in a sense, referring to Smith, the name, even on that occasion, did not. To repeat, a semantic ambiguity is out of the question in these cases, and it looks as if the familiar pragmatic machinery will handle the differences perspicuously.

In fact, Kripke uses his pair of name examples to state two difficulties for any proposed semantical version of the referential-attributive distinction. The first difficulty has already been discussed. It is based upon the following subtheses. (i) The pragmatic account is independently required to deal with simple and complex cases involving proper names, and (ii) this same account is also adequate to explicate the phenomena involved in "referential uses" of descriptions. From these, it is inferred that the pragmatic account is to be preferred. Now, (i) is true, of course, and this establishes that the existence of simple and complex cases with definite descriptions does not support a semantical distinction for descriptions. What does support such a distinction, from my perspective, is the more complicated set of considerations described in sections I and II. Correspondingly, these same considerations make the case against thesis (ii). It is not plausible that all deictic uses of pronominal descriptions can be treated in a satisfactory way by means of the pragmatic apparatus. Therefore, what Kripke here calls "the unitary account" is not as broadly adequate as these last remarks of his suggest.

But Kripke goes on to formulate a second puzzle. "Further, the resemblance between the case of descriptions and that of proper names (where presumably no one would be tempted to postulate an ambiguity) is so close that any attempt to explain the cases differently is automatically suspect" (*op. cit.*, p. 20). This passage can be construed as offering a challenge to the semantic version of the referential-attributive distinction. Since the phenomena in the pair of name cases are so closely analogous to the phenomena in the counterpart attributive and referential cases, it is extremely difficult to understand how it is that there is no semantic ambiguity in the name cases while, in the description cases, such an ambiguity supposedly exists. The challenge to the advocate of a semantic distinction for descriptions is to explain how this can possibly be.

Oddly enough, a crucial hint as to how this challenge should be met is found in another of Kripke's footnotes.

> There is one significant difference between the case of proper names and that of definite descriptions. If someone uses "Jones" to refer to Smith, he has *misidentified* Smith as Jones, taken Smith for someone else. To some extent I *did* think that *Jones* was raking the leaves. (I assume that "Jones" is already in his idiolect as a name of Jones. If I am introduced to an impostor and am told, "This man is none other than Albert Einstein," if I am fooled I will have *taken* him, falsely, to be Einstein. Someone else, who has never heard of Einstein before, may merely be mistaken as to the impostor's name.) On the other hand, if I think that someone is "her husband" and so refer to him, I need not at all have confused two people. I merely think that one person possesses a property—that of being married to her—that in fact he lacks. The real husband is irrelevant (*op. cit.*, p. 25, fn. 26).

The delicate issue that this passage raises but does not attempt to answer is: "In what sense is the actual husband 'irrelevant' to the statement?"

The answer that I have proposed, in effect, proceeds along the following lines. Consider the assertion

(16) Brown is her husband.

In one obvious and perfectly standard use, (16) means

Brown is a man who is (uniquely) married to her.

That is, the predicate 'is her husband' simply ascribes the indicated complex property or condition to Brown. In relation to the predicate, only this property is semantically (and pragmatically) relevant. If we imagine, for the sake of illustration, that (16) is false and that the lady is actually married to Smith, then (16) does not express, strictly and literally, the singular proposition

of Smith, that Brown is (identical with) *him*

Smith, the actual husband, is semantically (and pragmatically) irrelevant to the meaning that the predicate, on that occasion, expresses. Now, once this is allowed, there is no reason not to recognize that definite descriptions in other grammatical positions may function in the same way. Thus, there is a natural (attributive) reading of

(17) Her husband is kind to her.

which says (roughly)

A man [some man or other] who is (uniquely) married to her is kind to her.

In this utterance also, 'her husband' functions as a kind of quantifier phrase (i.e., 'The *x:* husband of her'), and the utterance does not even purport to say *of* Smith that he is kind to the woman. That is, the actual husband is once more—and in the very same way—semantically irrelevant. For the advocate of semantically ambiguous descriptions, this is the crucial difference that is *not* reflected in simple and complex cases with proper names. The words 'Jones is raking leaves', in either case, literally say of Jones that he is raking leaves—they literally express a singular proposition of which Jones is a constituent. But, by contrast, the semantical version of the referential-attributive distinction explicitly endorses the point suggested by Kripke's footnote. If the description in (17) is attributive, then the proposition literally expressed by the words is a *general* proposition (at least in relation to what the full description contributes).[18] It is only when the description is pronominal and used deictically that a genuine singular proposition is expressed. Therefore, since the purported contrast between the attributive and the referential, properly understood, is of a different character from the contrast between simple and complex uses of names, the absence of a semantical ambiguity in the names does not impugn the coherence and integrity of the semantical distinction for descriptions. To repeat a previous disclaimer, however, the evidence that this distinction is actually exemplified in our use of descriptions must depend on facts other than the mere existence of simple and complex cases that feature definite descriptions.

These contentions will be substantially clarified if we see that Kripke's pragmatic apparatus does indeed have a very general applicability. In particular, it has an important role to play *even within* a framework that adopts the semantical version of the referential-attributive distinction. The familiar pragmatic discriminations are required in cases where a description occurs attributively *and* in cases where a pronominal description is used deictically.

Here, first of all, is an "attributive" example typical of what I have in mind. Greg expresses to Brian his desire to earn $50 so that he will be able to buy a new bike, and he wonders how he might get the money. Brian recalls that the annual Cub Scout lightbulb sale is just underway and that the most successful Scout salesman wins a $50 prize. He also knows, as does Greg, that Greg has been the winner for two years running and that, given his superlative sales talents, Greg is sure to win again this year. So Brian, meaning to remind Greg of all this, says

> The Scout who sells the most lightbulbs this year will win $50.

Brian's use of the definite description is plainly attributive and expresses a thought about whomever sells the most bulbs. It is *not* a pronominal description making reference to Greg. In this context, however, Brian has intended to suggest to Greg that he (Greg) will shortly win a $50 prize. But this thought, the chief thought behind Brian's remark, is not something literally expressed by the words even on this occasion. It is merely a thought which Brian meant, by his utterance, to convey. This example follows the lines laid down by Kripke exactly.

---

[18] Wettstein has emphasized that this is what purports to be the crucial contrast in his article, "Demonstrative Reference and Definite Descriptions," *Philosophical Studies*, XL (1981): 241–57.

Moreover, it is my opinion that there are "referential" cases that also give rise to the same pragmatic effects. Suppose that a pair of people, *S* and *S**, know a certain man, Jones, reasonably well. Although he is a familiar acquaintance, they happen not to have learned his name and know him and standardly refer to him as "Muriel's husband." The phrase 'Muriel's husband' is, as it were, a description *of* Jones *for* *S* and *S**. (We also assume that Jones *is* Muriel's husband.) Finally, we add that *S* and *S** also know that Muriel's husband is notoriously lax as a gardener and almost never does any lawn work. The following exchange now takes place:

> *S*: (pointing into the distance) There is Muriel's husband over there.
> *S**: What is he doing?
> *S*: You'll never believe this. Muriel's husband is raking leaves.

This exchange can be situated in variant contexts in such a way that the cases thus generated will be either simple or complex. It is simple if the man in the distance is Jones, i.e., Muriel's husband. But, if the man in the distance is actually Smith, whom, perhaps because of the bad light, the speakers have mistakenly taken to be Jones, then the case is a complex one. (It will be obvious that this has been constructed in analogy to Kripke's "name" example.)

In either of these cases, on the view I have defended, *S*'s concluding words say literally of Jones—under the correct presupposition that he is Muriel's husband—that he is raking leaves. And, in the complex case, those same words do *not* say literally of Smith that he is raking leaves. When Smith is the man in the distance what *S*'s words express in the context is true if Jones, even in some other locale, is raking leaves. Naturally, in producing his utterance, *S* has, in some sense, said of Smith that he is raking leaves, but that proposition, as before, is simply a part of what *the speaker* meant on the given occasion. The pragmatic apparatus applies here just as it did when proper names were on the stage. It should be agreed on all sides that none of this motivates any further semantical distinctions. The meaning of an attributive description and the meaning of a pronominal description remains unaffected as we pass, in each instance, from simple to complex cases. And this is wholly consistent with what Kripke has showed us concerning the use of proper names.

Entwined in many of the issues considered here are questions about what notion of "singular reference" is taken to be crucial for semantics and about how the favored notion is to be construed. I shall conclude with some brief remarks on this elusive subject. When Kripke introduces the concept of "semantic reference," he stipulates that the semantic referent of a name is its bearer and that the semantic referent of a definite description is the object that uniquely satisfies the descriptor. Taken together, these stipulations imply that it is the semantic function of a description in normal contexts to refer or purport to refer to its unique satisfier. And it is presupposed that 'refers' applies unequivocally to descriptions and proper names. If, however, attributive definite descriptions are quantifier phrases, as I have suggested, then these assumptions may well be wrong. It seems clear in any case that Donnellan intended to reject them in passages like this: "If there is anything which might be identified as reference here [in attributive uses], it is reference in a very weak sense—namely, reference to *whatever* is the one and only one ϕ."[19]

It should be admitted that the import of this passage is cloudy, since the supposed "weak" and "strong" sense of 'refers' is not really delineated. Kripke, in his article, has some excellent remarks about the dangers of postulating an ambiguity when the only grounds for this are that the difference in senses is needed to save some conceptual claim (*op. cit.*, p. 19). One can easily wonder whether Donnellan is not guilty of such a move in the claim cited above, but, although the suspicion is reasonable, I think that there is a simple defense of Donnellan's intuitions.

The verb 'refers' in English is plainly ambiguous in a relevant way. For example, we say

> The phrase 'a stockbroker' *refers* to any person who professionally executes orders to buy and sell securities.

However much philosophers of language may deplore this usage, it is a common one, and its meaning is clear enough: it means "correctly applies to" or "is true of." Adapting some terminology from Russell, let us call "referring" in this sense "denoting." By contrast, if we have a statement with a pure demonstrative, e.g., This is ψ, then we say that the relevant token of 'this' *refers to* such-and-such an object. In making such a report, we do not mean that the word 'this' correctly applies to or is true of the object referred to. Indeed, it is rather difficult to say what exactly we do mean. Very roughly, the token of the demonstrative, construed in relation to other features of the utterance, serves to signal that the speaker's words are meant to say, literally and expressly, *of* some contextually indicated object that *it* is ψ. The pure demonstrative thus systematically contributes to the expression of what is meant to be a singular proposition that contains the referent of the demonstrative as a constituent. Appropriating Russell's terminology once more, let us call reference in this sense "designating."[20]

---

[19] "Reference and Definite Descriptions," p. 302. He seems to suggest that it is a mark of an attributive use of ⌜the ϕ⌝ that the relevant utterance can be paraphrased as ⌜. . . the ϕ, whoever he (she, it) is, . . .⌝. Certainly many subsequent writers have supposed that this possibility of appositive interjection provides a criterion of attributive use. The test is a bad one, however, and it has led to confusion. I can appropriately assert, while demonstrating a salient individual in front of me, "That man, whoever he is, is a loud buffoon," if I fail to possess whatever interest-relative identificatory information for which the situation and my audience call. Surely my appositive disclaimer does not make my use of the demonstrative 'nonreferential'. On the other hand, we may notice that 'whoever', 'whichever', etc., sometimes function as quantifiers, e.g., in Frege's curious example, "Whoever touches pitch defiles himself." It would be a better test of attributive use to ask whether an utterance of ⌜The ϕ is ψ⌝ can be paraphrased as ⌜Whatever is the ϕ is ψ⌝, where 'whatever' is a quantifier or part of a quantifier phrase. The present quote from Donnellan should be read in this manner.

With this distinction loosely drawn, let me restate what I take to be Donnellan's idea in the quotation set out above. A definite description used attributively refers or purports to refer to its unique satisfier only in the "denoting" sense of 'refers'. The description purports to apply correctly or to be true of at most one object. But, such a description does not designate or purport to designate either the unique satisfier or any other object. The relevant utterance expresses a general proposition, and, therefore, designation of an object by the description is not semantically in play. In the context of his article as a whole, Donnellan is urging that we easily and naturally confuse this point because of the double use of 'refers' and because the very same expression type can also function as a pronominal description used deictically. In the latter case, the description *does* purport to designate an item that satisfies the associated descriptor. I take this interpretation of Donnellan to be confirmed by the substantial emphasis he places on the concept of "saying such-and-such *of* some object" in his discussions of referential use. Thus, in a passage just prior to the one quoted above, he states:

> We saw that when a definite description is used referentially, a speaker can be reported as having said something *of* something. And in reporting what it was of which he said something we are not restricted to the description he used, or synonyms of it; we may ourselves refer to it using any descriptions, names, and so forth, that will do the job (*op. cit.*, p. 303).

One might wish that the passage said more plainly that it is not just the speaker, but also his utterance, which says something *of* something, but here and elsewhere in his article it appears that "referential use" is to be understood in terms of reference qua designation. Reference qua denotation is relevant principally as a source of confusion in discussions of the topic.

---

[20] I have chosen "denoting" and "designating" here because they are suggestive in relation to Russell's use of the terms. But what I say is not at all meant as an explication of *his* concepts.

Definite descriptions occupy a curiously schizophrenic position within the realm of "singular" or "referring" terms. In attributive use, they may refer to an object, but they refer only in the sense that other predicables do: they are true of the object. Descriptions that are narrowly referential (pronominal descriptions used deictically) also refer, but this is the "reference" of names and demonstratives. If ⌜the ϕ⌝ saturates ⌜is ψ⌝, the whole may be used to say of a contextually determined item o that it is ψ. It is little wonder that the "referential" schizophrenia now diagnosed, has produced a parallel schizophrenia in the history of the topic.

Gottlob Frege subsumed definite descriptions within his broader category of "proper names" and took their referents, if any, to be the unique object described. In fact, he tends implicitly to treat attributively used descriptions as paradigms of the class of proper names. This creates pressures, at least in the exposition of his doctrines, to assimilate the relation of a name to its bearer to the relation of a description to its unique satisfier. The consequent tensions in his system have become well-known. Thus, in a certain system-relative sense, definite descriptions *are* "referring expressions" for Frege. Russell, we noted, divides "reference" into denotation and what he calls "direct designation." The latter is a relation which (probably) demonstratives bear to the objects demonstrated and which "logically proper names" would have to their bearers if any names were ever logically proper. Descriptions, though they may be said to denote, never directly designate anything. So, in this special Russellian sense, descriptions are not "referring expressions" for him. Peter Strawson insisted again that they were. But, on the whole, he tends to veer between the basic Fregean model and a model that stresses deictic use without an accompanying gloss on the attributive features that Russell, in effect, had stressed. It was Donnellan who first saw that there was something radically wrong with the whole debate. He had the insight that each theory was both highlighting one use over the other while, at the same time, muddling properties of the different uses together. Although Donnellan did not bring out the full generality of the linguistic phenomena he observed and was self-admittedly unsure about the semantical status of the two uses, his insight is central for the theory of reference and should not, as it might, be lost.

GEORGE M. WILSON

Johns Hopkins University

NEW BOOKS

ACHINSTEIN, PETER: *Particles and Waves: Historical Essays in the Philosophy of Science*. New York: Oxford, 1991. 337 p. Cloth $49.95, paper $24.95.

ACKERMANN, ROBERT JOHN: *Nietzsche: A Frenzied Look*. Amherst: Massachusetts UP, 1990. xi, 205 p. Cloth $24.95.

ADDELSON, KATHRYN PYNE: *Impure Thoughts: Essays on Philosophy, Feminism, and Ethics*. Philadelphia: Temple, 1991. vi, 257 p. Cloth $27.95.

ALLISON, HENRY E.: *Kant's Theory of Freedom*. New York: Cambridge, 1990. xii, 304 p. Cloth $49.50, paper $15.95.

ATWELL, JOHN E.: *Schopenhauer: The Human Character*. Philadelphia: Temple, 1990. xi, 259 p. Cloth $39.95.

BAIER, ANNETTE C.: *A Progress of Sentiments: Reflections on Hume's* Treatise. Cambridge: Harvard, 1991. xi, 333 p. Cloth $39.95.

BARTKY, SANDRA LEE: *Feminism and Domination: Studies in the Phenomenology of Oppression*. New York: Routledge, 1990. xiii, 141 p. Paper $12.95.

BARWISE, JON, and JOHN ETCHEMENDY: *The Language of First-Order Logic: Including the Program* Tarski's World. Stanford: CSLI, 1990. xiii, 259 p. Paper $27.50 (with software).

BERLEANT, ARNOLD: *Art and Engagement*. Philadelphia: Temple, 1991. xvii, 259 p. Cloth $49.95.

BERNADETE, SETH: *The Rhetoric of Morality and Philosophy: Plato's* Gorgias *and* Phaedrus. Chicago: University Press, 1991. ix, 205 p. Cloth $29.95.

BRAMS, STEVEN J.: *Negotiation Games: Applying Game Theory to Bargaining and Arbitration*. New York: Routledge, 1990. xviii, 297 p. Cloth $49.50, paper $15.95.

BROADIE, SARAH: *Ethics with Aristotle*. New York: Oxford, 1991. xiii, 462 p. Cloth $55.00.

BURNYEAT, MYLES: *The* Theaetetus *of Plato*. With a translation by M. J. LEVETT. Indianapolis: Hackett, 1990. xiv, 351 p. Cloth $32.50, paper $10.95.

CARR, DAVID: *Educating the Virtues: An Essay on the Philosophical Psychology of Moral Development and Education*. New York: Routledge, 1991. x, 294 p. Cloth $55.00.

CARRIER, DAVID: *Principles of Art History Writing*. University Park: Pennsylvania State UP, 1991. xiii, 249 p. Cloth $29.95.

CAVELL, STANLEY: *Conditions Handsome and Unhandsome: The Constitution of Emersonian Perfectionism*. Chicago: University Press, 1990. xxxix, 151 p. Paper $8.95.

COLLINS, H. M.: *Artificial Experts: Social Knowledge and Intelligent Machines*. Cambridge: MIT, 1990. xiii, 266 p. Cloth $19.95.

COOPER, DAVID E.: *Existentialism: A Reconstruction*. Cambridge: Blackwell, 1990. ix, 201 p. Cloth $52.95, paper $17.95.

COTHEY, A. L.: *The Nature of Art*. New York: Routledge, 1990. xi, 201 p. Cloth $74.50.

CRAIG, EDWARD: *Knowledge and the State of Nature: An Essay in Conceptual Synthesis.* New York: Oxford, 1990. x, 171 p. Cloth $49.95.

CRIMMINS, JAMES E.: *Secular Utilitarianism: Social Science and the Critique of Religion in the Thought of Jeremy Bentham.* New York: Oxford, 1990. xi, 348 p. Cloth $68.00.

DAVIES, STEPHEN: *Definitions of Art.* Ithaca: Cornell, 1991. x, 243 p. Cloth $31.50, paper $12.95.

DOUBLE, RICHARD: *The Non-Reality of Free Will.* New York: Oxford, 1990. xi, 247 p. Cloth $32.50.

DREYFUS, HUBERT L.: *Being in the World: A Commentary on Heidegger's* Being and Time, Division I. Cambridge: MIT, 1991. xiv, 370 p. Cloth $30.00, paper $15.95.

DRYZEK, JOHN S.: *Discursive Democracy: Politics, Policy, and Political Science.* New York: Cambridge, 1990. x, 254 p. Cloth $39.50.

DUMMETT, MICHAEL: *The Logical Basis of Metaphysics.* Cambridge: Harvard, 1991. xi, 355 p. Cloth $34.95.

EAGLETON, TERRY: *Ideology: An Introduction.* New York: Verso, 1991. xv, 242 p. Cloth $59.95, paper $17.95.

ECO, UMBERTO: *The Limits of Interpretation.* Bloomington: Indiana UP, 1990. 295 p. Cloth $27.50.

FENVES, PETER D.: *A Peculiar Fate: Metaphysics and World-History in Kant.* Ithaca: Cornell, 1991. xii, 306 p. Cloth $36.50.

FEREJOHN, MICHAEL: *The Origins of Aristotelian Science.* New Haven: Yale, 1991. ix, 174 p. Cloth $24.50.

FETZER, JAMES H.: *Philosophy and Cognitive Science.* New York: Paragon, 1991. xvii, 170 p. Paper $13.95.

FEYERABEND, PAUL: *Three Dialogues on Knowledge.* Cambridge: Blackwell, 1991. 167 p. Cloth $42.95, paper $17.95.

FLANAGAN, OWEN: *Varieties of Moral Personality: Ethics and Psychological Realism.* Cambridge: Harvard, 1991. xiv, 393 p. Cloth $34.95.

FRONGIA, GUIDO, and BRIAN MCGUINNESS: *Wittgenstein: A Bibliographical Guide.* Cambridge: Blackwell, 1990. x, 438 p. Cloth $65.00.

GELVEN, MICHAEL: *Truth and Existence: A Philosophical Inquiry.* University Park: Pennsylvania State UP, 1990. ix, 204 p. Cloth $25.00.

GORSHOLZ, EMILY R.: *Cartesian Method and the Problem of Reduction.* New York: Oxford, 1991. ix, 161 p. Cloth $42.50.

GOSLING, JUSTIN: *Weakness of the Will.* New York: Routledge, 1990. ix, 221 p. Cloth $39.95.

HACKER, P. M. S.: *Wittgenstein: Meaning and Mind: Volume 3 of an Analytic Commentary on the* Philosophical Investigations. Cambridge: Blackwell, 1990. xxi, 575 p. Cloth $95.00.

HACKING, IAN: *The Taming of Chance.* New York: Cambridge, 1990. xiii, 264 p. Cloth $44.50, paper $14.95.

HATFIELD, GARY: *The Natural and the Normative: Theories of Spatial Perception from Kant to Helmholtz.* Cambridge: MIT, 1990. xii, 366 p. Cloth $35.00.

HELLER, AGNES: *A Philosophy of Morals.* Cambridge: Blackwell, 1990. xiv, 245 p. Cloth $39.95. And *Can Modernity Survive?* Berkeley: California UP, 1990. vi, 177 p. Cloth $25.00.

HELLER, MARK: *The Ontology of Physical Objects: Four-Dimensional Hunks of Matter*. New York: Cambridge, 1990. xiv, 162 p. Cloth $34.50.

HONDERICH, TED: *Conservatism*. Boulder: Westview, 1990. ix, 255 p. Cloth $45.00.

HOOKWAY, CHRISTOPHER: *Scepticism*. New York: Routledge, 1990. xii, 251 p. Cloth $49.95.

HUNT, LESTER H.: *Nietzsche and the Origin of Virtue*. New York: Routledge, 1991. xxiii, 200 p. Cloth $49.95.

HYLTON, PETER: *Russell, Idealism, and the Emergence of Analytic Philosophy*. New York: Oxford, 1990. xvii, 420 p. Cloth $59.00.

IHDE, DON: *Instrumental Realism: The Interface between Philosophy of Science and Philosophy of Technology*. Bloomington: Indiana UP, 1991. xiv, 159 p. Cloth $29.95, paper $12.50.

INGRAM, DAVID: *Critical Theory and Philosophy*. New York: Paragon, 1990. xxx, 240 p. Paper $12.95.

JACKENDOFF, RAY: *Semantic Structures*. Cambridge: MIT, 1990. xiv, 322 p. Cloth $34.95.

JAMESON, FREDRIC: *Postmodernism, or, the Cultural Logic of Late Capitalism*. Durham: Duke, 1990. xxii, 438 p. Cloth $34.95.

KATZ, JERROLD: *The Metaphysics of Meaning*. Cambridge: MIT, 1990. xiii, 356 p. Cloth $29.95.

KEMP, ANTHONY: *The Estrangement of the Past: A Study in the Origins of Modern Historical Consciousness*. New York: Oxford, 1991. xi, 228 p. Cloth $32.50.

KITCHER, PATRICIA: *Kant's Transcendental Psychology*. New York: Oxford, 1990. xiii, 296 p. Cloth $35.00.

KLEIN, MARTHA: *Determinism, Blameworthiness, and Deprivation*. New York: Oxford, 1990. vi, 232 p. Cloth $39.95.

KRAYNAK, ROBERT P.: *History and Modernity in the Thought of Thomas Hobbes*. Ithaca: Cornell, 1990. vii, 224 p. Cloth $28.50.

KUKATHAS, CHANDRAN, and PHILIP PETIT: *Rawls:* A Theory of Justice *and Its Critics*. Stanford: University Press, 1990. xi, 169 p. Cloth $22.50, paper $8.95.

KYBURG, HENRY E.: *Science and Reason*. New York: Oxford, 1990. x, 283 p. Cloth $35.00.

LAUDAN, LARRY: *Science and Relativism: Some Key Controversies in the Philosophy of Science*. Chicago: University Press, 1990. xiii, 180 p. Cloth $32.00, paper $12.95.

LEHRER, KEITH: *Metamind*. New York: Oxford, 1990. viii, 313 p. Cloth $39.95.

LEVINSON, JERROLD: *Music, Art, & Metaphysics: Essays in Philosophical Aesthetics*. Ithaca: Cornell, 1990. xi, 419 p. Cloth $49.95, paper $16.95.

LEWIS, DAVID: *Parts of Classes*. Appendix by JOHN P. BURGESS, A. P. HAZEN, and DAVID LEWIS. Cambridge: Blackwell, 1991. x, 155 p. Cloth $39.95, paper $12.95.

LIPTON, PETER: *Inference to the Best Explanation.* New York: Routledge, 1991. x, 194 p. Cloth $49.95.

LLOYD, G. E. R.: *Demystifying Mentalities.* New York: Cambridge, 1990. viii, 174 p. Cloth $49.50, paper $14.95.

MACINTYRE, ALASDAIR: *Three Rival Versions of Moral Enquiry: Encyclopedia, Genealogy, and Tradition.* Notre Dame: University Press, 1990. x, 241 p. Cloth $24.95.

MCGINN, COLIN: *The Problem of Consciousness: Essays towards a Resolution.* Cambridge: Blackwell, 1991. viii, 216 p. Cloth $47.95.

MCINERNEY, PETER K.: *Time and Experience.* Philadelphia: Temple, 1991. vii, 283 p. Cloth $39.95.

MARTIN, MICHAEL: *The Case against Christianity.* Philadelphia: Temple, 1991. xi, 273 p. Cloth $44.95.

MEINWALD, CONSTANCE C.: *Plato's* Parmenides. New York: Oxford, 1991. vii, 192 p. Cloth $32.50.

MICHALSON, GORDON E.: *Fallen Freedom: Kant on Radical Evil and Moral Regeneration.* New York: Cambridge, 1990. xii, 172 p. Cloth $34.50.

MISAK, C. J.: *Truth and the End of Inquiry: A Peircean Account of Truth.* New York: Oxford, 1991. xiii, 182 p. Cloth $49.95.

MORTON, ADAM: *Disasters and Dilemmas: Strategies for Real-life Decision Making.* Cambridge: Blackwell, 1991. xii, 209 p. Cloth $49.95.

NEUHOUSER, FREDERICK: *Fichte's Theory of Subjectivity.* New York: Cambridge, 1990. x, 180 p. Cloth $39.50, paper $12.95.

NUSSBAUM, MARTHA C.: *Love's Knowledge: Essays on Philosophy and Literature.* New York: Oxford, 1990. xix, 403 p. Cloth $42.50.

PARSONS, TERENCE: *Events in the Semantics of English: A Study in Subatomic Semantics.* Cambridge: MIT, 1990. xviii, 334 p. Cloth $35.00.

PEARS, DAVID: *Hume's System: An Examination of the First Book of His Treatise.* New York: Oxford, 1990. ix, 204 p. Cloth $49.95.

POWELL, C. THOMAS: *Kant's Theory of Self-Consciousness.* New York: Oxford, 1990. xi, 268 p. Cloth $56.00.

PREVOST, ROBERT: *Probability and Theistic Explanation.* New York: Oxford, 1990. v, 195 p. Cloth $59.00.

PRIOR, WILLIAM J.: *Virtue and Knowledge: An Introduction to Ancient Greek Ethics.* New York: Routledge, 1991. xi, 240 p. Cloth $59.95, paper $15.95.

ROSENFELD, MICHEL: *Affirmative Action and Justice: A Philosophical and Constitutional Inquiry.* New Haven: Yale, 1991. ix, 373 p. Cloth $30.00.

RUBEN, DAVID-HILLEL: *Explaining Explanation.* New York: Routledge, 1990. xi, 265 p. Cloth $39.95.

SCHAEFER, DAVID LEWIS: *The Political Philosophy of Montaigne.* Ithaca: Cornell, 1990. xv, 407 p. Cloth $41.50, paper $13.95.

SCHEFFLER, ISRAEL: *In Praise of the Cognitive Emotions and Other Essays in the Philosophy of Education.* New York: Routledge, 1991. xii, 174 p. Cloth $39.95, paper $13.95.

STATEN, HENRY: *Nietzsche's Voice*. Ithaca: Cornell, 1990. xii, 223 p. Cloth $29.95

STERELNY, KIM: *The Representational Theory of Mind: An Introduction*. Cambridge: Blackwell, 1990. xiii, 252 p. Cloth $44.95, paper $16.95.

STOVE, DAVID: *The Plato Cult and Other Philosophical Follies*. Cambridge: Blackwell, 1991. xiii, 209 p. Cloth $26.95.

THOMSON, JUDITH JARVIS: *The Realm of Rights*. Cambridge: Harvard, 1990. viii, 383 p. Cloth $35.00.

TITO, JOHANNA MARIA: *Logic in the Husserlian Context*. Evanston: Northwestern, 1990. 291 p. Cloth $36.95.

TORRETTI, ROBERTO: *Creative Understanding: Philosophical Reflections on Physics*. Chicago: University Press, 1990. xvi, 369 p. Cloth $79.95, paper $32.50.

TURBAYNE, COLIN MURRAY: *Metaphors of the Mind: The Creative Mind and Its Origins*. Columbia: South Carolina UP, 1991. xv, 133 p. Cloth $24.95.

VAN INWAGEN, PETER: *Material Beings*. Ithaca: Cornell, 1990. viii, 299 p. Cloth $32.50.

WALLER, BRUCE N.: *Freedom without Responsibility*. Philadelphia: Temple, 1990. x, 236 p. Cloth $39.95.

WARDY, ROBERT: *The Chain of Change: A Study of Aristotle's* Physics VII. New York: Cambridge, 1990. x, 345 p. Cloth $59.50.

WEST, CORNEL: *The Ethical Dimensions of Marxist Thought*. New York: Monthly Review, 1991. xxxiv, 183 p. Cloth $36.00, paper $18.00.

WHITE, ALAN: *Within Nietzsche's Labyrinth*. New York: Routledge, 1990. xv, 188 p. Cloth $39.95, paper $13.95.

WILLIAMSON, TIMOTHY: *Identity and Discrimination*. Cambridge: Blackwell, 1990. ix, 173 p. Cloth $45.95.

WOLF, SUSAN: *Freedom within Reason*. New York: Oxford, 1990. xii, 162 p. Cloth $24.95.

WOLFF, JONATHAN: *Robert Nozick: Property, Justice, and the Minimal State*. Stanford: University Press, 1991. xiii, 168 p. Cloth $32.50, paper $9.95.

WOLIN, RICHARD: *The Politics of Being: The Political Thought of Martin Heidegger*. New York: Columbia, 1990. xvii, 221 p. Cloth $29.50.

WOOD, ALLEN W.: *Hegel's Ethical Thought*. New York: Cambridge, 1990. xxi, 293 p. Cloth $49.50.

WYSCHOGROD, EDITH: *Saints and Postmodernism: Revisioning Moral Philosophy*. Chicago: University Press, 1990. xxvii, 298 p. Cloth $40.00, paper $15.95.

YOUNG, IRIS MARION: *Justice and the Politics of Difference*. Princeton: University Press, 1990. viii, 286 p. Cloth $45.00, paper $12.95. And *Throwing Like a Girl and Other Essays in Feminist Philosophy and Social Theory*. Bloomington: Indiana UP, 1990. vii, 213 p. Cloth $29.95, paper $14.50.

THE JOURNAL OF PHILOSOPHY

VOLUME LXXXVIII, NO. 8, AUGUST 1991

HOLISM AND NONSEPARABILITY*

Metaphysics follows physics in fact as well as in Aristotle. This is hardly a novel insight as applied to the views of great philosophers of the past, such as Descartes, Leibniz, or Kant. But it holds morals for contemporary thought also. One moral is that much contemporary metaphysics, rather than proceed independently of any particular view of the physical world, in fact rests on uncritical assumptions as to its nature which are hard to reconcile with contemporary physics. A second moral is that radical developments in physics, such as the theories of relativity and quantum mechanics, are themselves a rich source of metaphysical insights, irrespective of the ultimate fate of those theories or of particular interpretations of them.

The claim that quantum phenomena exhibit a characteristic holism or nonseparability provides an important illustration of both morals. Niels Bohr and Erwin Schrödinger put forward such ideas in the early days of quantum theory, but they have been repeated more insistently in recent years by a number of physicists and philosophers.[1] Metaphysicians cannot afford to ignore these ideas. But this

---

* For constructive comments on an earlier version of this paper, I thank Michael Jubien, Paul Teller, Martin Jones, and other members of our Northern California discussion group, as well as the philosophy department at the University of Arizona.

[1] Bohr often remarked upon the holistic character of quantum descriptions. He stressed the fact that such descriptions are significant only in the context of an entire experimental arrangement, which he took to be required for the meaningful attribution of properties to its elements, including the atomic objects on which it may be used to perform measurements. Schrödinger, on the other hand, emphasized the idea of nonseparability. He stressed the fact that individual quantum systems often fail to have separate quantum states of their own, even though the compound quantum system they compose does have such a state. See Bohr, *Atomic Theory and the Description of Nature* (New York: Cambridge, 1934); Schrödinger, "Discussion of Probability Relations between Separated Systems," *Proceedings of the Cambridge Philosophical Society*, XXXI (1935): 555–63.

0022-362X/91/8808/393–421

does not mean that claims of holism or nonseparability should be accepted without subjecting them to detailed critical analysis. It is the purpose of this paper to contribute to such an analysis. Only when we know just what the theses of holism and nonseparability amount to will we be in a position to assess the extent to which quantum phenomena establish these theses and to explore their further metaphysical implications.

This task is made more difficult but at the same time more important by many philosophers' natural suspicions of holism. Such suspicions may even be strengthened by a survey of previous attempts to characterize quantum holism.[2] Clean and interesting holist theses need to be formulated if the concept is to be rehabilitated.

I shall offer formulations of both holism and nonseparability. These turn out to be interestingly different, but closely related notions. Indeed, they are so closely related that it is not surprising that previous analyses have not adequately distinguished them. Whether quantum mechanics establishes either holism or nonseparability depends on how one interprets quantum mechanics. But the metaphysician should not use controversy about the correct interpretation of quantum mechanics as an excuse for not investigating the implications of holism and nonseparability; for quantum mechanics comes about as close to demonstrating holism and nonseparability as it does to establishing indeterminism. Just as in the case of indeterminism, the prospect of holism and nonseparability presents a stimulus, not a threat, to metaphysics.

I

Holism is often characterized as the view that the whole is more than the sum of its parts. This has a paradoxical ring to it, since on one natural understanding it is obviously false: any whole consists precisely of its parts, no more and no less. If the whole seemed to be more than the sum of its parts, then that could only be because additional parts were being overlooked. Nevertheless, it will serve as a suggestive metaphor guiding the search for more acceptable formulations of holism.

---

[2] This is particularly true of Bohr's notoriously obscure discussion of holism. But more recent discussions of holism and nonseparability have been more suggestive than precise. See, among others, David Bohm, *Wholeness and the Implicate Order* (New York: Routledge & Kegan Paul, 1980); Bernard d'Espagnat, *In Search of Reality* (New York: Springer, 1983); Don Howard, "Einstein on Locality and Separability," *Studies in History and Philosophy of Science*, xvi (1985): 171–201, and "Holism, Separability and the Metaphysical Implications of the Bell Experiments," in J. T. Cushing and E. McMullin, eds., *Philosophical Consequences of Quantum Theory: Reflections on Bell's Theorem* (Notre Dame: University Press, 1989); Michael Redhead, *Incompleteness, Nonlocality and Realism: A Prolegomenon to the Philosophy of Quantum Mechanics* (New York: Oxford, 1987). Paul Teller has proposed what he calls "relational holism" as an explication of the kind of holism or nonseparability evidenced in the quantum domain. His analyses may be fruitfully compared with those I offer here: see, in particular, his "Relational Holism and Quantum Mechanics," *British Journal for the Philosophy of Science*, xxxvii (1986): 71–81, and "Relativity, Relational Holism, and the Bell Inequalities," in J. T. Cushing and E. McMullin, eds. The present paper owes much to the work of these and other recent authors; I have tried to straighten, smooth, and signpost the trail they have blazed.

As a first step, consider the nature of the composition relation. If one were to take this to be fully captured by the axioms of mereology, then to an *arbitrary* subset of objects S from a given set there would correspond a whole $M(S)$—the mereological sum of S. If none of the objects in S overlaps another, the objects in S may be said to compose $M(S)$. Note that, with composition understood in terms of mereology, the objects in S compose the whole $M(S)$ irrespective of their properties or relations. On one natural understanding, holism would then become trivially false; for $M(S)$ is precisely the (mereological) sum of its parts.

We can try to rescue the thesis of holism from the fate of trivial falsity by noting that for something to be a typical whole of a certain kind it is not enough for the mereological sum of its parts to exist. For there to be a model airplane is not just for the mereological sum of its parts to exist at a time, even if these do then compose it. If the builder had left the kit unopened the previous day instead of constructing the model, the mereological sum of its parts would have existed at that time without thereby composing this (or indeed any) airplane. Other examples show that the properties of the individual parts are just as important as their spatial relations to their composing a whole of a certain kind. A particular object would not have been the nuclear bomb that it is if the atoms of uranium 235 which it contains had had the same half-life (4.5 billion years) as atoms of uranium 238.

Clearly many composite objects are in some sense more than just mereological sums of their parts. But in what sense? Call any part of an object which is neither identical to that object, nor a part of every object, a *proper* part. An object is *composite* just in case it has proper parts. Two proper parts *overlap* if one is a part of the other or if there is an object that is a proper part of both of them. A set of proper parts of an object no two of which overlap is a *nonoverlapping set*. A nonoverlapping set P of proper parts of an object o such that every part of o overlaps some element of P *composes* o: the objects in P are then *component parts* of o (relative to the partition P, though the qualification may often be dropped without confusion). We may call a composite object o a *genuine whole with component parts* P if P composes o, but the existence of each element of P is not a sufficient condition for the existence of o.

There is then a trivial sense in which every genuine whole is more than the sum of its parts; for any genuine whole is distinct from the mereological sum of its components. All that is required for the existence of the mereological sum is the existence of the elements of *P*. But this does not suffice for the existence of the genuine whole. For the elements of *P* compose *o* only if each has the required properties, and these elements are appropriately arranged. This is true despite the fact that *o* is composed solely of elements of *P*.

Of course, one may still question whether there are any genuine wholes, as I have defined them. If, for example, one believes that an ordinary object like a model airplane, a table, or a nuclear bomb is identical to a certain mereological sum, then one will not accept that such an object is a genuine whole.[3] If there are no ordinary genuine wholes, one may begin to doubt whether there are any extraordinary ones either. But if in fact there are no genuine wholes, the fact that any genuine whole is more than the sum of its parts becomes even more trivially true!

How can we arrive at a formulation of holism that is neither trivially false nor trivially true? I think the key is to focus on respects in which the features of a whole may be independent of those of its proper parts.

Wholes frequently exhibit features that seem qualitatively distinct from those of their component parts (relative to some partition). The water in my kettle is wet and boils at 100 degrees Celsius, even though none of its component molecules is wet or has a boiling point. And copper is a red metal that conducts electricity, even though none of the protons, neutrons, or electrons that make it up is red or metallic or an electrical conductor. Properties of a whole possessed by none of its component parts (relative to some chosen partition) could be called emergent. One way of understanding holism is as the thesis that a whole of type *W* composed of a set *P* of proper parts is more than the sum of these parts in the sense that any object of type *W* has some property *F* while none of these component parts has property *F*. On this way of understanding it, a compound object could instantiate holism to various degrees, depending firstly on the choice of partition, and secondly on the variety and importance of emergent properties possessed by objects of that type.

---

[3] Jubien, for example, believes this; cf. his (unpublished) manuscript "Ordinary Naming and Necessity." He argues there that, even though an object such as a particular table is identical to a mereological sum, that mereological sum has certain properties, including the property of "being this table," which it might have lacked. On this view, moreover, the object with the property of "being this table" might have been a different mereological sum. Jubien takes this last consequence of his view to be the correct rendering of the intuition that this table might have had different parts from those which it in fact has. He is thus able to acknowledge a sense in which a certain *table* could have had different parts, even though the mereological sum that is that table could not. In general, on this view, an object's being a particular whole of a certain kind is not just a matter of its being a certain mereological sum, even though it *is* a mereological sum (and consequently *not* a genuine whole in my sense): for a different mereological sum could have uniquely instantiated the property of being the particular whole of that kind.

But holism in this sense is surely too familiar a feature of the world to occasion much surprise or comment. The attempt to recover all the sensible properties of physical objects starting from a severely restricted set of properties of their constituents has been a familiar theme in metaphysics since the pre-Socratics and in physical science at least since the time of Galileo. In fact, it would be more surprising if such attempts could be shown to be subject to serious obstacles. And this presents a clue to more interesting holist theses. Rather than assert that a whole has features that none of its components has, holism should be understood to claim that the whole has features that cannot be reduced to features of its component parts. But now we need to get clearer on what we mean by 'features' and 'reduction'.

There are at least two different approaches we can take at this point. The first approach leads to formulations of holism best suited for a discussion of explanatory or methodological issues, while the second leads to formulations that are of more direct relevance to metaphysics. Since my interest here is metaphysical, I shall be pursuing the second approach in what follows. But it will help to make this approach clearer if I begin by saying a few words about the kind of holism with which it is to be contrasted. For explanatory holism is perhaps a more familiar thesis, and while it is important to make clear that this is not the kind of holism with which I shall be primarily concerned, the truth of metaphysical holism would have interesting implications for explanatory holism. Epistemology can be divorced from metaphysics no more here than elsewhere.

II

Holism claims that a whole has features that cannot be reduced to features of its parts. If the features of an object include its behavior, and reducing the features of a whole to those of its parts means explaining its behavior in terms of their behavior, then we arrive at a thesis of explanatory holism. Explanatory holism is the view that a satisfactory explanation of the behavior of an object of some type cannot be given by analyzing that object into its component parts and appealing to the laws that hold of these parts. Thus, a biological holist may maintain that it is impossible fully to explain the functioning of an entire organism purely in terms of the physical and chemical laws obeyed by the molecules of which its cells are composed. And a social holist may maintain that an adequate explanation of the behavior of a society will require more than knowledge of the principles governing the actions of the society's individual members.[4]

Explanatory holism thus stands opposed to explanatory reductionism—the view that one can achieve a completely adequate understanding of the behavior of the whole by analyzing it into its component parts and discovering the laws according to which these operate. Such a reductive methodology has proven enormously successful in physical science since the time of Galileo, so that today it is common to assume that elementary particle physics lies at the foundation of all physical science, and that, if only physicists knew the laws governing the fundamental constituents of matter, they would truly have "a theory of everything." Nevertheless, there continue to be influential dissenting voices, even among physicists. Thus in his recent book, *The Problems of Physics*, A.J. Leggett[5] considers the physics of condensed matter. He writes

> . . . there are also a number of cases, particularly in the traditional areas of the physics of gases and crystalline solids, in which a model which treats the behavior of the whole as essentially just the sum of that of its parts (atoms or electrons) has been quite successful. . . . But these cases, despite the fact that they totally dominate the presentation of the subject in most elementary textbooks, are actually the exception rather than the rule. In virtually all the frontier areas of modern condensed-matter physics, the relationship between our understanding of the behavior of matter at the microscopic level of single atoms and electrons, and at the macroscopic level of (say) liquids and solids, is actually a good deal more complicated than this. . . . I would claim that the most important advances in this area come about by the emergence of qualitatively new concepts at the intermediate or macroscopic levels—concepts which, one hopes, will be compatible with one's information about the microscopic constituents, but which are in no sense logically dependent on it (*ibid.*, p. 113).

Even within physics, then, there continues to be debate about the respective merits of explanatory reductionism and explanatory holism. If metaphysical holism is true, then there is at least one domain in which reductive explanations will be unavailable. This may prompt a reassessment of research strategy also in other domains. If so, metaphysics may come to influence methodology.

---

[4] Of course, other varieties of biological and social holism may prove more worthy of debate.

[5] New York: Oxford, 1987.

But just what are the theses of explanatory reductionism and explanatory holism? A somewhat more precise formulation requires preliminary discussion of ultimate and basic parts. An ultimate part of an object is a part that itself has no proper parts. Whether or not an object has ultimate parts is a question for physics, and not just for philosophy. Atoms are no longer considered ultimate constituents of matter by contemporary physics. Although electrons and quarks are better candidates for that title, the discovery of particles at successively more elementary levels below the level of the atom has caused some physicists to question whether matter *has* ultimate constituents, and even if it does, whether we shall ever be in a position to know what they are. But whether or not matter has ultimate parts, we may still wish to speak of the basic parts of an object. For example, even though atoms are not the ultimate constituents of matter, there is still a sense in which they may be claimed to be the basic components of ordinary matter. But in what sense? We may try to capture that sense as follows:

A set *B* of objects is *basic* relative to a domain *D* of objects and processes of type *P* just in case

(1) Every element of *D* is composed solely of objects in *B*,
(2) the objects in *B* do not overlap,
(3) while processes of type *P* may create or destroy elements of *D*, or may convert one element of *D* into another, no process in *P* can create, destroy, or interconvert elements of *B*, and
(4) there is no set *C* satisfying (1)–(3) such that some element of *C* has some element of *B* as a proper part.

Intuitively, the basic parts relative to *D* and *P* may be thought of as the *building blocks* of elements of *D*: the largest components that retain their integrity through processes in *P* capable of building, destroying, and transforming elements of *D*. A *basic part* of an object *o* (relative to a domain *D* and processes *P*) is just a part of *o* which is basic relative to *D* and *P*.

It is a consequence of these definitions that some basic parts may be ultimate, while others may not be. Ultimate parts of an object are basic relative to any domain of objects that includes it and to any processes that neither create, destroy, nor exchange these ultimate parts. But note that the so-called elementary particles will not be basic parts of all objects relative to all processes; for there are processes that create, destroy, and interconvert elementary particles.

We may now formulate the following theses:

Pure physical explanatory reductionism: The physical behavior of any domain *D* of physical objects subject only to processes *P* can be explained in terms of the physical properties and relations of their basic physical parts (relative to *D* and *P*).

> *Pure physical explanatory holism:* There is some domain *D* of physical objects subject only to processes *P*, not all of whose physical behavior can be explained in terms of the physical properties and relations of their basic physical parts (relative to *D* and *P*).

Each of these theses is described as "pure" because each is restricted in its application to physical objects. One might formulate a more ambitious thesis of physical explanatory reductionism according to which the behavior of *every* object can be thus explained. This more ambitious thesis would form a natural supplement to the physicalist view that all objects are, at bottom, physical objects. The pure explanatory theses, on the other hand, are naturally understood to presuppose the following more restricted principle:

> *Pure physical composition:* Every basic part of a physical object (relative to some domain *D* and processes *P*) is itself a physical object.

III

The kind of holism I want to focus on now may be called metaphysical rather than explanatory. A thesis of metaphysical holism stands opposed to a claim of supervenience in some domain. The supervenience claim is that all properties of individuals in the domain are determined by the properties of their parts.[6] 'Partism' would be a suggestive, but ugly, name for this claim: in the interests of harmonious usage, I adopt Teller's term 'particularism' instead. But note that a proponent of particularism in this sense need bear no animosity toward universals. It is rather metaphysical holism that the particularist is concerned to deny. The metaphysical holist claims that some individuals in the domain in question have properties that are not determined by the properties of their parts. We need to see how to make these claims more precise in the case of physical holism.

The first point to note is that, if we are to arrive at an interesting formulation of metaphysical holism, we must accept that this thesis is not only concerned with properties, and not concerned with all properties. The properties of a genuine whole will typically depend upon *relations* among its proper parts as well as on properties of the individual parts. But if we are permitted to consider *all* properties and relations among the parts, then these trivially determine the properties of the whole they compose. For one relation among the parts is what we might call the *complete composition relation*—that relation among the parts which holds just in case they compose this very whole with all its properties.

---

[6] While the notion of supervenience is itself not exactly unproblematic (as a large recent philosophical literature attests), an analysis of holism in terms of supervenience may still contribute to philosophical progress. I take the determination involved here to be not logical but physical (in some suitably broad sense).

Let us call the set of properties and relations of the parts upon which the properties and relations of the whole may or may not supervene the *supervenience basis*. To avoid trivializing the theses we are trying to formulate, only certain properties and relations can be allowed in the supervenience basis. The intuition as to which these are is simple—the supervenience basis is to include just the qualitative, intrinsic properties and relations of the parts, i.e., the properties and relations that these bear in and of themselves, without regard to any other objects, and irrespective of any further consequences of their bearing these properties for the properties of any wholes they might compose. Unfortunately, this simple intuition resists precise formulation. It is notoriously difficult to say precisely what is meant either by an intrinsic property or relation, or by a purely qualitative property or relation.[7] And the other notions appealed to in expressing the simple intuition are hardly less problematic. But, imprecise as it is, this statement serves already to exclude certain properties and relations from the supervenience basis.

In particular, besides excluding the complete composition relation, it excludes what we might call historical properties and relations, and individual properties and relations. A historical property of an object is a property that object has at a time simply in virtue of what happens to it at some other time. Having been driven 30,000 miles is a historical property of my car. An individual property of an object is a property whose instantiation depends upon the existence of some individual. It is both a historical and an individual property of the Mona Lisa that it was painted by Leonardo da Vinci.

The only properties and relations of the parts which are allowed in the supervenience basis are those which the relevant parts actually manifest at the time in question, regardless of any other individuals, and irrespective of what happens to the parts at other times. Notice that some but not all properties and relations in the supervenience basis may be essential properties of the relevant parts—properties or relations those parts must have if they are to exist at all.

Now it is clear that historical or individual properties of a whole cannot be expected to supervene on a basis of properties and relations of their parts which fails to include their historical or individual properties and relations. Consequently, the twin claims of particularism and metaphysical holism should be restricted so that each applies only to the qualitative, intrinsic properties of the whole. On the other hand, these claims may be broadened so that they apply not merely to a single whole, but also to a set of wholes of some kind. If the claims are so broadened, then the determination of any relations among the elements of the set by the properties and relations of their parts in the supervenience basis is just as much in question as the determination of any of their individual properties.

---

[7] David Lewis explicitly discusses the notion of an intrinsic property, and implicitly also that of an intrinsic relation, in *On the Plurality of Worlds* (New York: Blackwell, 1986). He argues there that appeal to a basic unanalyzed notion of intrinsic property is philosophically legitimate, while also pointing to two alternative attempts to explicate that notion. In his earlier paper, "Extrinsic Properties," in *Philosophical Studies*, XLIV (1983): 197–200, he criticized another, simpler attempted explication of this notion. A property or relation may be said to be qualitative if it does not depend on the existence of any specific individual. But this does not, of course, constitute a satisfactory definition.

Finally, we arrive at the following opposing theses:

> *Pure physical particularism:* Every qualitative, intrinsic physical property and relation of a set of physical objects from any domain *D* subject only to processes of type *P* is supervenient upon the qualitative, intrinsic physical properties and relations of their basic physical parts (relative to *D* and *P*).
>
> *Pure physical holism:* There is some set of physical objects from a domain *D* subject only to processes of type *P*, not all of whose qualitative, intrinsic physical properties and relations are supervenient upon the qualitative, intrinsic physical properties and relations of their basic physical parts (relative to *D* and *P*).

As in the case of the theses of explanatory holism and reductionism considered earlier, each of these theses may be taken to presuppose pure physical composition.[8]

It is easy to give examples that conform to pure physical particularism. The temperature of some particular sample of gas supervenes upon the translational kinetic energies of its constituent molecules. My body's currently being attracted toward the center of the earth with a force of some 160 pounds weight supervenes upon the masses of each part of my body and of the earth, and on the distance between each part of my body and each part of the earth.

But one cannot provide familiar illustrations of pure physical holism, precisely because exceptions to pure physical particularism lie outside the range of our normal physical experience. This is what gives quantum holism metaphysical bite. It is worth considering why it is so difficult to come up with such illustrations. I think the main reason is that physical composition is naturally thought of as resulting from spatial juxtaposition. Even though there are examples of intrinsic relations that are not spatial, such examples seem to depend on intrinsic spatial relations. The size of the electrostatic attraction between two charged particles is an intrinsic relation between the particles which is not a spatial relation. It does, however, depend on the spatial distance between the particles as well as on their individual electric charges. Genuine physical wholes seem built up out of their parts by putting these parts together in the right spatial relations.

---

[8] One might try to explicate the notion of supervenience here in terms of models of a true, descriptively complete, physical theory. At issue is whether there exist two models of such a physical theory of some objects in *D* which agree on the qualitative, intrinsic physical properties and relations of the basic parts of these objects but disagree on some qualitative intrinsic property or relation of the objects themselves.

There are many cases in which this turns out to be true, even when at first sight other causal relations between the parts seem to be playing an independent role in the constitution of the whole. It may seem that all the parts of an amplifier are correctly spatially arranged even though it does not work properly, but closer examination may reveal a short circuit—itself a product of inappropriate spatial relations among the small scale parts of the amplifier. Frequently, intrinsic relations among parts of a genuine whole which do not supervene on intrinsic properties and spatial relations of *those* parts do supervene on intrinsic properties and spatial relations of more elementary parts. We have a strong intuition, based on a great deal of physical experience, that the intrinsic physical properties and relations of composite objects are determined by the intrinsic physical properties and spatial relations of their component parts, even though the parts for which this is true may be microscopic or even ultimate.

It might be objected that the reason it is so difficult to come up with illustrations of physical holism is that there can be none, because this thesis is a priori false. I can think of two arguments for this conclusion, but neither is successful. Here is the first argument. Let w be a particular physical object composed of a set $P(w)$ of basic physical parts, which have the necessary, intrinsic physical properties and relations to compose w. Suppose that w has some physical property J even though the $P(w)$ might have had all the same intrinsic physical properties and relations without its being true that w has J. Now, the $P(w)$ are physically arranged precisely so as to compose w, which has J. Hence the set $P(w)$ of basic physical parts has the necessary intrinsic physical properties and relations to compose w, complete with the property J. Thus J is in fact supervenient upon the intrinsic properties and relations of the $P(w)$, and so pure physical particularism holds after all.

This argument embodies a modal fallacy. It is fallacious to infer from the premise that the $P(w)$ has all the necessary intrinsic physical properties and relations to compose w, which has J, to the conclusion that these properties and relations are incompatible with w's failing to have *J*. Only if *J* were an essential property of *w* would *w*'s existence guarantee that it has *J*. But not all intrinsic properties are essential. To illustrate pure physical holism, all that is necessary is to find some nonessential intrinsic physical property of a whole which does not supervene on the intrinsic physical properties and relations of its physical parts.

The second argument is not so easy to dispose. Suppose, as before, that *w* is composed of physical parts *P*(*w*), and that *w* in fact has some nonessential intrinsic property *J*. There is a relation *R* among the *P*(*w*) which can be defined as follows: *R* holds among the *P*(*w*) if and only if *w* exists and has *J*. Moreover, this relation *R* should be included in the supervenience basis along with any other qualitative, intrinsic physical relations among the *P*(*w*). But if *R* is included in the supervenience basis, then this is after all sufficiently complete to determine not only that *w* exists but also that *w* has *J*. Hence pure physical particularism holds here after all.

The response to this argument is to question the assumption that the relation *R* should be included in the supervenience basis. Recall that the complete composition relation *C* was earlier disqualified from appearing in the supervenience basis on the grounds that *C* is not a qualitative, intrinsic physical relation among the *P*(*w*)—a relation that holds without regard to any further consequences of its holding for any whole that the *P*(*w*) might compose. The relation *R* should also be excluded from the supervenience basis on these same grounds. It is not clear that the relation *R* is a physical relation. Even if *R* is counted as a physical relation, it seems clear that it is not an *intrinsic* physical relation, even though the property *J* to which it corresponds is an intrinsic physical property of *w*. *R* is not a relation that the *P*(*w*) have in and of themselves: it is a relation they inherit only insofar as they compose a whole *w* with property *J*.[9]

---

[9] In fact, more needs to be said in response to the second argument. For certainly some nonessential, qualitative intrinsic physical properties of some wholes do supervene on the qualitative, intrinsic physical relations as well as properties of their component parts. If *J* is such an intrinsic property of a whole *w*, then there exists a qualitative, intrinsic physical relation *R*′ among the *P*(*w*) whose holding helps to determine that *w* has *J*. Moreover, there seems no a priori justification for excluding cases in which the holding of *R*′ among the *P*(*w*) is both necessary and sufficient for *w* to exist and have *J*. But in such a case, there is a relation (*R*′) in the supervenience basis which, of necessity, holds among the *P*(*w*) if and only if *w* exists and has *J*. And this seems to show that the relation *R* defined earlier cannot, after all, be automatically excluded from the supervenience basis.

Even in such a case, however, the relations *R* and *R*′ are not identical, and while *R*′ must be included in the supervenience basis, *R* should not be. For *R* it is a matter of definition that it holds among the *P*(*w*) if and only if *w* exists and has *J*, while for *R*′ this is not true by definition, but only as a matter of what might be called natural necessity. In terms of the metaphor of possible worlds, *R* and *R*′

Although this response to the second argument seems intuitively powerful, it is not as conclusive as the response to the first argument, because the distinctions on which it rests have not been given precise analyses. Even in the absence of such analyses, the response may be made more persuasive by exhibiting a concrete case of a whole w, a property J, and relation R among the basic parts $P(w)$ of w in which it is intuitively clear that, while J is an intrinsic physical property of w, it is equally clear that R is not an intrinsic physical relation among $P(w)$. The a priori arguments against pure physical holism would look much weaker in the face of what intuition recognizes as a concrete case of such holism. That is what quantum mechanics may be able to provide, as will be discussed in section VI below.

IV

Although philosophers are familiar with discussions of holism, the notion of nonseparability may require some introduction. In fact, there is at least one philosopher who has seen the metaphysical import of this notion, namely, David Lewis, who I believe would reject it. For the thesis of Humean supervenience that he advances in the preface to his *Philosophical Papers, Volume II*[10] comes close to formulating the contrary notion of separability. Humean supervenience is the doctrine that

> . . . all there is to the world is a vast mosaic of local matters of particular fact, just one little thing and then another. (But it is no part of the thesis that these local matters are mental.) We have geometry: a system of external relations of spatiotemporal distance between points. Maybe points of spacetime itself, maybe point-sized bits of matter or aether or fields, maybe both. And at those points we have local qualities: perfectly natural intrinsic properties which need nothing bigger than a point at which to be instantiated. For short: we have an arrangement of

---

agree in their extensions in the actual world and in all naturally possible worlds, but not in *all* possible worlds. They are therefore different relations. It is a remarkable natural fact that R' holds among the $P(w)$ if and only if w exists and has J, whereas it is a trivial consequence of its definition that R holds among the $P(w)$ if and only if w exists and has J. Indeed, the former fact would be *so* remarkable as to be extremely implausible. It would be very surprising for any whole w to exist and have such an intrinsic property J irrespective of the intrinsic properties of the $P(w)$. Even when an intrinsic physical relation is included in the supervenience basis for some whole, in no ordinary case does that relation exhaust this basis. Consequently, in no ordinary case is there an intrinsic physical relation in the supervenience basis whose extension is included in that of R even in the actual world: coincidence in extension in all naturally possible worlds would be quite extraordinary. The cases we have been considering in the present paragraph are therefore quite bizarre. It is satisfying that the exclusion of R from the supervenience basis can be given a sound justification even in such bizarre cases.

[10] New York: Oxford, 1986.

> qualities. And that is all. There is no difference without difference in the arrangement of qualities. All else supervenes on that (*ibid.*, p. x).

If one restricts the scope of Humean supervenience so that it claims just that all purely *physical* phenomena supervene on perfectly natural intrinsic *physical* properties that need nothing bigger than a point at which to be instantiated, then one naturally arrives at a physical principle that may be formulated in terms of the notion of spatiotemporal separability. A physical process may be said to be *spatiotemporally separable* in *R* (where *R* is a space-time region) if and only if it is supervenient upon an assignment of qualitative, intrinsic physical properties at space-time points in *R*.

> *Spatiotemporal separability:* Any physical process occurring in a region *R* of space-time is spatiotemporally separable in *R*.

A physical process that violates the principle of spatiotemporal separability in some region may be said to be *nonseparable*. *Nonseparability* holds just in case some physical process is nonseparable. If spatiotemporal separability holds, then every physical process is determined by what happens locally—that is to say, by what is happening at each space-time point where that process is going on.

It is a natural assumption that most, if not all, ordinary physical processes are spatiotemporally separable. For example, the boiling of a kettle of water is spatiotemporally separable in a region containing the kettle over a period of time during which the water boils. The boiling of the water consists in the increased kinetic energy of its constituent molecules permitting each to overcome the short range attractive forces that otherwise hold it in the liquid. It thus supervenes on the assignment, at each space-time point on the trajectory of each molecule, of intrinsic physical properties to that molecule (such as its kinetic energy), perhaps together with intrinsic physical properties representing the magnitude and direction of the attractive force acting on that molecule at that point. The fall of a stone is similarly spatiotemporally separable in a space-time region that contains it. For the stone consists of atoms, each of which has a space-time trajectory on which various intrinsic mechanical properties are defined at each point, and each atom is acted on by various force fields, importantly including the gravitational field due to the earth, which are also defined at each point on its trajectory. It is worth noting, however, that spatiotemporal separability would appear to fail even here if one attempted to describe the falling stone without assigning a gravitational field at points of space-time. For such a description would require a specification of the distance of the stone from the center of the earth, and this is not an intrinsic physical property (or relation) holding at points confined to the region in which the stone is falling. We see here that maintaining the separability principle provides a motive for the introduction of fields, where these are taken to involve assignment of intrinsic physical properties to, or at, points of space-time independently of whether or not these points are occupied by enduring physical objects.

More generally, any physical process described fully by a local space-time theory will conform to the principle of spatiotemporal separability. For such a theory proceeds by assigning geometric objects (such as vectors or tensors) to each point in space-time to represent physical fields, and then requiring that these satisfy certain field equations.[11] But processes described fully by theories of other forms will also be spatiotemporally separable. This is true not only of pure field theories, but also of many theories that assign properties to particles at each point on their trajectories. Of familiar physical theories, it is only theories involving direct action between spatially separated particles whose models are nonseparable, and even they describe processes that are spatiotemporally separable within sufficiently large regions of space-time. It is interesting that the reformulation or replacement of such theories by field theories has often been thought of as a conceptual advance. I conjecture that concern to preserve the principle of spatiotemporal separability is one significant reason behind this perception (while another, more familiar, reason is provided by the wish to exclude action-at-a-distance).

No one has been more strongly drawn to field theories in physics than Albert Einstein. It is therefore not surprising that, in his writings on quantum mechanics, he himself formulated principles closely related to the principle of spatiotemporal separability. Immediately after offering one such formulation, he relates it to field theory.

> If one asks what, irrespective of quantum mechanics, is characteristic of the world of ideas of physics, one is first of all struck by the following: the concepts of physics relate to a real outside world, ... It is further characteristic of these physical objects that they are thought of as arranged in a space-time continuum. *An essential aspect of this arrangement of things in physics is that they lay claim, at a certain time, to an existence independent of one another, provided these objects 'are situated in different parts of space'.* . . . This principle has

---

[11] The notion of a local space-time theory is defined in John Earman and John Norton, "What Price Spacetime Substantivalism? The Hole Story," *British Journal for the Philosophy of Science*, XXXVIII (1987): 515–25. It is implicit in Michael Friedman's *Foundations of Spacetime Theories* (Princeton: University Press, 1983), which formulates special and general relativity, as well as Newtonian kinematics and gravitational theory, as local space-time theories.

> been carried to extremes in the field theory by localising the elementary objects on which it is based and which exist independently of one another, as well as the elementary laws which have been postulated for it, in the infinitely small (four-dimensional) elements of space.[12]
>
> (I have emphasized the sentence in this passage which, following Howard, I take to express Einstein's separability principle.)

It may appear at first sight that the principle Einstein formulates here is concerned with spatial rather than spatiotemporal separability. But the fact that Einstein goes on to refer to four-dimensional elements of *space*, together with the fact that the phrase 'are situated in different parts of space' appears in quotes, seem to indicate that his principle is indeed to be understood as concerning spatiotemporal separation. In the next section, I shall consider another principle which may be abstracted from Einstein's writings which is concerned explicitly with *spatial* rather than spatiotemporal separability.

V

Holism and nonseparability are quite distinct notions, as is now apparent. Holism has to do with the irreducibility of certain part-whole relations, while nonseparability is to be understood in spatiotemporal terms. Why, then, have philosophers of physics confused the two?[13] This may well be because there are such subtle and interesting connections between these notions. It is these which I now propose to explore.

Recall the following statement of pure physical holism:

> *Pure physical holism:* There is some set of physical objects from a domain *D* subject only to processes of type *P*, not all of whose qualitative, intrinsic physical properties and relations are supervenient upon the qualitative, intrinsic physical properties and relations of their basic physical parts (relative to *D* and *P*).

There are two related respects in which this formulation remains problematic. The notion of an intrinsic physical relation is purely formal: it is unclear what would count as such a relation. But in the absence of such clarification, the second a priori argument against pure physical holism still has weight. Without any clear idea as to what would count as an intrinsic physical relation, how can one justify excluding the relation R defined in section III from that category? The following reformulation solves both problems, but at a price:

---

[12] This passage is quoted from pp. 170–1 of *The Born-Einstein Letters* (New York: Macmillan, 1971). It is a translation by Irene Born from the original German included in a letter of Einstein to Born dated 5 April 1948, in which Einstein encloses a copy of an essay that appeared as "Quantenmechanik und Wirklichkeit," in *Dialectica*, II (1948): 320–4.

[13] See, for example, Howard, "Holism, Separability and the Metaphysical Implications of the Bell Experiments," and Teller, "Relativity, Relational Holism, and the Bell Inequalities," both in J.T. Cushing and E. McMullin, eds.; and Redhead.

> *Pure spatial holism:* There is some set of physical objects from a domain D subject only to processes of type P, not all of whose qualitative, intrinsic physical properties and relations are supervenient upon the qualitative, intrinsic physical properties and the spatial relations of their basic physical parts (relative to D and P).

Instead of speaking of intrinsic physical relations among basic parts, pure spatial holism mentions only their spatial relations. Spatial relations are of special significance because they seem to yield the only clear examples of qualitative, intrinsic physical relations required in the supervenience basis in addition to the qualitative, intrinsic properties of the relata.[14] Other intrinsic relations supervene on spatial relations. This is true, for example, of the physical relation that holds between two point masses just in case they are presently attracting one another by a force of one gram weight (at least according to Newtonian gravitational theory): for this relation supervenes on the masses and their spatial separation. By restating the principle of physical holism so as to refer only to spatial relations in specifying the supervenience basis, one renders that principle more concrete, apparently without affecting its essential thrust. Moreover, since the relation R defined in section III is clearly not a spatial relation, one cannot argue that pure spatial holism is trivially false on the grounds that R must be included in that basis because it is an intrinsic physical relation. But the resulting formulation is less elegant because it treats intrinsic properties and relations differently. And it is less general, because it fails to leave open the possibility that the progress of physics may reveal additional nonsupervenient, intrinsic physical relations that should also be included in the supervenience basis.

The shift from pure physical holism to pure spatial holism has introduced some spatiotemporal relations, and has thereby linked a holist thesis more closely to nonseparability. We can forge a connecting link in the other end of the chain by introducing the notion of spatial separability.

---

[14] Teller in fact believes that spatial relations themselves supervene on intrinsic physical properties of objects (see his paper, "Space-time as a Physical Quantity," in R. Kargon and P. Achinstein, eds., *Kelvin's Baltimore Lectures and Modern Theoretical Physics* (Cambridge: MIT, 1987), pp. 425–47). If this turns out to be correct, then the reference to spatial relations in the statement of pure spatial holism will prove to be redundant.

Consider some physical process involving interactions among a number of enduring physical systems throughout some interval of time: for example, the motions of the planets around the sun. There will be some compound system (such as the solar system) of which these are all component subsystems. If the process is spatiotemporally separable, then it is supervenient upon an assignment of qualitative, intrinsic physical properties at space-time points in the region in which it occurs. Suppose that qualitative, intrinsic property assignments are made at just those space-time points within the region which lie on the trajectory of each component subsystem. (Such an assignment would suffice to specify the dynamical behavior of the solar system.) If all these assignments, together with the spatial relations obtaining between the components at each moment, are compatible with each of two or more different assignments of qualitative, intrinsic physical properties to the compound system at each moment throughout the interval, then the process is nonseparable. This kind of nonseparability results from a breakdown of what might be called *spatial separability*.

> *Spatial separability:* The qualitative, intrinsic physical properties of a compound system are supervenient on the qualitative, intrinsic physical properties of its spatially separated component systems together with the spatial relations among these component systems.

A number of Einstein's formulations of separability occur in the context of his discussions of pairs of separated systems, such as those which occur in the famous Einstein-Podolsky-Rosen paradox. Here is one such formulation:

> The real state of [the pair AB] consists precisely of the real state of A and the real state of B, which states have nothing to do with one another.[15]

I think we can recognize the spatial separability principle as a natural generalization of Einstein's formulation, if we read the last clause of the latter, not as excluding causal influence, but rather as requiring that the real states of *A* and *B* not be derivative from some more fundamental state of the compound system *AB*.[16] It is also possible to see the spatial separability principle as a generalization and correction of the principle Howard states in his article, "Einstein on Locality and Separability" (see fn. 2/3), that "any two spatially separated systems possess their own separate real states" (*op. cit.*, p. 173). Correction is necessary since Howard's principle would hold if two spatially separated systems had separate real states, even when the whole they composed had a real state that did *not* supervene on the separate real states of its components.

---

[15] This formulation appears in a letter from Einstein to Schrödinger dated June 19, 1935. I have quoted it from Howard, "Einstein on Locality and Separability," p. 180. The Einstein-Podolsky-Rosen paradox itself first appears in Einstein, B. Podolsky, and N. Rosen, "Can Quantum-Mechanical Description of Physical Reality be Considered Complete?" *Physical Review*, XLVII (1935): 777–80.

[16] In fact, Einstein himself may not have understood the principle this way. For he takes the following locality condition (which appears in his next sentence) to follow immediately from his spatial separability principle, which on my reading it does not, without additional causal assumptions. "The real state of B . . . cannot

Spatiotemporal separability entails spatial separability. The argument is as follows. Suppose that spatial separability fails. Then there is a time t at which some compound system has a qualitative, intrinsic physical property P that does not supervene upon the qualitative intrinsic properties and relations of its spatially separated components. If some of these components are localized at spatial points, then their qualitative, intrinsic physical properties will generate assignments of qualitative, intrinsic physical properties at the space-time points defined by these spatial points at t. There may be other ways in which the qualitative, intrinsic physical properties and relations of the components restrict the compatible assignments of qualitative, intrinsic physical properties at space-time points occurring at t. Let R_t be the space-time region defined by a time-slice at t. Consider an assignment A of qualitative, intrinsic physical properties at space-time points in R_t which is compatible with the qualitative, intrinsic physical properties and relations of the components at t. If spatiotemporal separability held, then a process involving the compound system at t would have to be separable in R_t. But, by assumption, there is such a process that involves the compound system's having property P at t. Hence assignment A would have to determine that the compound system has P at t. But A was any assignment compatible with the qualitative, intrinsic physical properties and relations of the components at t. And so the qualitative, intrinsic physical properties and relations of the components at t would have to determine that the compound system has P at t. This would contradict the earlier conclusion that P does not supervene upon the quali-

---

depend on the kind of measurement I carry out on A." Note that, if one takes the real state of a system to be given by its quantum-state vector (as does one popular version of the Copenhagen interpretation), then quantum mechanics immediately conflicts with Einstein's spatial separability principle. For the singlet quantum state of a system AB composed of two subsystems A and B, each of nonzero spin, does not consist of the quantum state of A and the quantum state of B. But Einstein, for one, did *not* take the real state of a system to be given by its quantum-state vector. Indeed, one of his strongest reasons for denying that the quantum state of a system gives its real state, and consequently for rejecting this version of the Copenhagen interpretation, was his desire to interpret quantum mechanics in a way consistent with his spatial separability principle.

tative, intrinsic physical properties and relations of its components at *t*. Therefore, spatiotemporal separability fails. Contrapositively, spatiotemporal separability entails spatial separability.

Spatial separability is therefore linked to spatiotemporal separability because it is entailed by it. But spatial separability is also closely connected to spatial holism. To make this clear, consider the following denial of spatial separability:

> *Spatial nonseparability:* There exists a compound physical system, not all of whose qualitative, intrinsic physical properties supervene on the qualitative, intrinsic physical properties of its spatially separated component systems together with the spatial relations among these component systems.

Compare this with the principle of spatial holism stated earlier.

> *Pure spatial holism:* There is some set of physical objects from a domain *D* subject only to processes of type *P*, not all of whose qualitative, intrinsic physical properties and relations are supervenient upon the qualitative, intrinsic physical properties and the spatial relations of their basic physical parts (relative to *D* and *P*).

These two principles are clearly very similar, even though I have taken one to express a form of holism, and the other to express a form of nonseparability. With inchoate versions of these last two principles in mind, it would be easy to blur the distinction between holism and nonseparability, and to regard these distinct notions as essentially equivalent.

On one natural supposition, pure spatial holism entails spatial nonseparability. For suppose that pure spatial holism obtains because there is a qualitative, intrinsic physical property *P* of some compound system which does not supervene on the qualitative, intrinsic physical properties and spatial relations of its basic physical parts at some time when these parts are all spatially separated. In that case it follows that *P* also fails to supervene on the qualitative, intrinsic physical properties of these spatially separated component systems together with the spatial relations among them, and so spatial nonseparability obtains. (It might not be enough for the basic parts merely to *compose* the spatially separated components, for if they did so in such a way that the properties of any spatially separated component failed to supervene on the properties of its basic parts, then pure physical holism might hold while spatial nonseparability failed.)

If spatiotemporal separability entails spatial separability, then spatial nonseparability entails nonseparability. And if pure spatial holism also entails spatial nonseparability, it then follows that pure spatial holism entails nonseparability. Moreover, provided that spatial relations are included among qualitative, intrinsic physical relations, pure physical holism entails pure spatial holism, in which case it further follows that pure physical holism entails nonseparability. The foregoing arguments license one to detach the consequents of these conditionals subject only to the (rather weak) provisos these arguments required.

But it is quite clear that nonseparability does not entail pure physical holism. A process may be nonseparable even though it involves only objects with no nontrivial components (i.e., objects that are elementary parts of any object that contains them). The passage of an electron through the famous two-slit experiment may well provide an instance of such a process; for there is no reason to suppose that an electron has nonzero spatial extension, or any nontrivial component parts, whereas there is some reason to maintain that an electron passes through both slits of this experiment at once, and so has a nonseparable trajectory.[17] Perhaps, then, quantum mechanics furnishes examples of nonseparability without holism. In the final section, I consider whether it provides examples of nonseparability arising from an underlying holism.

VI

It follows from the foregoing analysis that, in order to determine whether or not quantum mechanics establishes either pure physical holism or nonseparability, it is necessary first to decide what are the qualitative, intrinsic physical properties and relations of quantum systems. This is by no means a straightforward matter. It may be easier to begin with the corresponding question about classical systems.

Classical particle mechanics ascribes two kinds of physical properties to the systems with which it deals. On the one hand, there are properties (such as a particular value of mass or charge) possessed by every system of particles of a given type, which remain unchanged as long as that system exists: we may call these the *fixed properties* of the system. On the other hand, there are properties associated with quantities called *dynamical variables* (such as energy; component of position, of momentum, of angular momentum). A system of a given type may have one of these *dynamical properties* at one time while failing to have it at some other time. For any "nice" set of real numbers, and any dynamical variable pertaining to a system, there is a dynamical property that the system possesses just in case the value of that variable on the system is restricted to that set. In particular, for each unit set, there is a dynamical property that the system has just in case the variable has the value contained in that set: we may call this a *precise value property*.

---

[17] The key feature of such examples is the attribution to a quantum system of nonlocalized spatial properties, such as an interval-valued (rather than a real-valued) position; cf. my *The Philosophy of Quantum Mechanics: An Interactive Interpretation* (New York: Cambridge, 1989), which offers an interpretation of quantum mechanics with this feature. See also Teller, "Quantum Mechanics and the Nature of Continuous Physical Quantities," this JOURNAL, LXXVI, 7 (July 1979): 345–61.

The possible dynamical properties of a system are generally taken to be determined by the structure of an abstract mathematical space (a phase-space) associated with systems of that type. The phase-space for a single particle is six-dimensional: each point in the space represents a precise real value for each of the three components of the particle's position, and each of the three components of its momentum. The phase-space for a compound system composed of N particles is $6N$–dimensional, and each point represents the position and momentum of every component particle. In classical particle mechanics, the instantaneous dynamical state of a system (whether simple or compound) is given by the assignment of a precise value property to each dynamical variable. It may be completely specified by assigning to the system a point in phase-space. Since all dynamical variables are functions of position and momentum, this assignment generates an ascription of a precise real value to every dynamical variable pertaining to that system. Each dynamical property of a system is associated with a "nice" subset of phase-space: the system has that property if and only if its dynamical state is represented by a point lying in that subset of phase-space.

It would be natural to assume that all fixed properties of a classical particle system are among its qualitative, intrinsic physical properties, as are all dynamical properties corresponding to any "nice" region of phase-space. One might also assume that these comprise the qualitative, intrinsic physical properties of a classical particle system, and that such systems have no qualitative, intrinsic physical relations that do not supervene on these. But both these assumptions may be questioned.

Among the dynamical properties of a classical particle are precise value properties associated with components of position. A property such as $x = 3$ ascribing a precise x-coordinate to the particle is not itself an intrinsic property of the particle, however. Rather, it corresponds to a spatial relation between that particle and an (arbitrarily chosen) origin of spatial coordinates, namely, the relation of being 3 units (in some system) distant from the origin along a certain spatial direction. That relation is actually an intrinsic relation involving the particle: its other relata may be particles, points of space or space-time, or a combination of these. Moreover, it does not seem to supervene on intrinsic physical properties of its relata.[18] If it is necessary to give an intrinsic specification of the dynamical state of the particle which implies the ascription to it of a property such as $x = 3$, then intrinsic spatial relations involving that particle must be included in the specification. But one might wonder whether that is necessary. For a one particle system is highly idealized, and one might suspect that, for a multiparticle system, the only spatial properties that need to be included in an intrinsic specification of its dynamical state are those connected to the spatial relations among just these particles.

Consider then a two-particle classical system. One dynamical property represented by a region of the 2-particle phase-space corresponds to the distance between the two particles being exactly 3 centimeters. As well as being a property of the 2-particle system, however, this also corresponds to a relation between the two particles that compose that system. Moreover, this is a qualitative, intrinsic physical relation between these two particles which may well not supervene on their qualitative, intrinsic physical properties. The phase-space representation of the state of a compound system does not represent qualitative, intrinsic physical relations among that system's components as such, and may therefore mislead one into overlooking some such relations. But in classical mechanics one need not employ this representation. It would be more perspicuous for present purposes to specify the dynamical state of a set of classical particles by giving their relative positions, in addition to their individual momenta, at each instant.[19]

Quantum mechanics also ascribes to its systems two kinds of qualitative, intrinsic physical properties, many of which are familiar from classical particle mechanics.[20] There are fixed properties, now also including intrinsic spin. And there are dynamical properties, including components of spin. In quantum mechanics, the dynamical properties of a system, whether simple or compound, are associated not with subsets of a phase-space, but with subspaces of a different kind of abstract mathematical space (a Hilbert space) appropriate to systems of that type.

---

[18] But cf. Teller, who proposes in his “Space-time as a Physical Quantity” that such spatial relations are indeed supervenient on qualitative, intrinsic physical properties of objects.

[19] One might think it unnecessary to give the momenta in addition, on the grounds that momentum is derivative on mass and velocity, and relative velocity may be derived from change in relative position. But classical dynamics also requires a specification of absolute acceleration that cannot be derived from relative velocities or their changes. This raises the interesting question as to whether further qualitative, intrinsic physical relations underlie the specification of momenta—perhaps including a relation that holds between a particle and points of space-time just in case that particle's trajectory is inertial.

[20] I restrict attention to the nonrelativistic quantum mechanics of a fixed set of particles. The case of quantum field theories raises further difficulties concerning

The instantaneous state of a system is not given in quantum mechanics by assigning a precise real value to every dynamical variable pertaining to that system. Rather, the state of a system at a time may be described by assigning to it a mathematical object of a certain kind—a wave function or, more generally, a *state vector* (which is an element of the Hilbert space associated with the system). For a system described by a given state vector at a particular time, quantum mechanics contains an algorithm that specifies, for each dynamical property, the probability that an observation of that property would reveal it to be possessed at that time. The algorithm is such that this probability is one just in case the state vector lies in the subspace associated with that property. It follows that there is no state vector that assigns probability one to some precise value property associated with each dynamical variable. In particular, no state vector assigns probability one to a precise value property associated with a given component of position and also to a precise value property associated with the same component of momentum.

What is the connection between the state vector of a system at a time and the dynamical properties it possesses at that time? Different interpretations of quantum mechanics answer this question differently. According to one popular version of the Copenhagen interpretation, a system has a dynamical property at a time if and only if it may be described by a state vector that lies in the subspace associated with that property. It is an immediate consequence of this version that no system ever has simultaneous precise values of both position and momentum. But on other interpretations one cannot "read off" a system's dynamical properties from its state vector in this way: the specification of a system's dynamical properties requires additional, or other, information. Call the set of dynamical properties possessed by a system at a time its *dynamical state* at that time. While interpretations differ as to whether or not a system's dynamical state is determined by a state vector associated with the

---

the character of the systems these theories may be taken to describe. In order to decide whether or not such theories manifest holism or nonseparability it will therefore be necessary, first, to resolve disputed ontological questions raised by quantum field theory.

system, on no interpretation does quantum mechanics itself prescribe the dynamical state except via this state vector.[21]

What, then, are the qualitative, intrinsic physical properties and relations of systems described by quantum mechanics? These clearly include a system's fixed properties. Presumably, they also correspond in some way to its dynamical properties, whether or not these are determined by any state vector. Perhaps a system's state vector should also be included among its qualitative, intrinsic physical properties. Quantum systems may even have further qualitative, intrinsic physical properties or relations that quantum mechanics does not describe at all (corresponding to so-called “hidden variables”). A decision as to whether or not quantum mechanics demonstrates nonseparability or physical holism requires a definitive characterization of the qualitative intrinsic properties and relations of quantum systems. But such a characterization cannot be given without resolving problems central to the interpretation of quantum mechanics. Nevertheless, one can show that there are significant interpretations of quantum mechanics according to which this theory does indeed demonstrate nonseparability, and arguably also pure physical holism.[22] Perhaps the easiest way of doing this requires a preliminary discussion of an important class of compound quantum systems, namely, spin-singlet pairs.

A pair of quantum systems *A*, *B*, each with spin ½, may be in a spin-singlet state. They then constitute a compound quantum system *AB* with certain properties, such as having total spin zero; and each of *A*, *B* separately is a basic part of *AB* relative to the domain *D* of pairs composed of subsystems of these types, and any type *P* of process through which systems of these types are preserved. (If, for example, *A* is an electron and *B* is a positron, then each may well be an ultimate part of *AB*, and in any case each is a basic part of *AB* relative to processes that do not create or annihilate electrons or positrons.) As long as *AB* remains in the spin-singlet state, it retains the dynamical property of having total spin zero. One fixed property of *AB* is that its electric charge is the sum of the electric charges of *A* and of *B*. Another property of *AB* is that it is described by a state vector that specifies a characteristic set of probabilistic correlations for joint measurements of spin components, one on *A*, the other on *B*. If, for example, the spin components along some one direction of *A* and of *B* are simultaneously measured, the probability that each has the same value is zero, while the probability that the two spin components have opposite values is one. Collectively, the spin-singlet correlations violate the famous Bell inequalities—a fact of some significance, since it rules out certain potential explanations of these correlations, as well as certain potential interpretations of quantum mechanics.[23]

---

[21] It is sometimes convenient to describe a system's state not by a vector, but by a more general mathematical object (e.g., a ray, or a density operator). Such descriptions are generally understood to be either equivalent to, or less detailed than, descriptions in terms of a state vector. While there are theories that provide a more detailed specification of the dynamical state of a system, these theories do not claim that quantum mechanics itself yields this specification. Rather, they supplement the quantum mechanical description, typically by the addition of so-called “hidden variables” whose values are not determined by the state vector. Bohm presents one such theory in his “A Suggested Interpretation of the Quantum Theory in Terms of ‘Hidden Variables’, I and II,” *Physical Review*, LXXXV (1952): 166–93.

[22] Since my main aim in this paper is to clarify, not to settle, the issues of quantum holism and nonseparability, it would be inappropriate for me to assume or defend any particular interpretation of quantum mechanics here. In my *The Philosophy of Quantum Mechanics: An Interactive Interpretation*, I propose and defend an interpretation that is similar in presently relevant respects to that version of the Copenhagen interpretation whose bearing on holism and nonseparability I consider below.

On the version of the Copenhagen interpretation mentioned above, a spin-singlet pair *A*, *B* clearly establishes nonseparability and arguably also pure physical holism. *A*, *B* are the basic parts of the compound system *AB*. On this view, the qualitative, intrinsic physical properties and relations of *A* correspond to just those fixed and dynamical properties recognized by quantum mechanics, and similarly for *B*. Since neither *A* nor *B* has a state vector in the singlet state, this cannot be included as a further intrinsic property of either.

Just as in classical particle mechanics, the instantaneous dynamical properties of *A*, *B*, and *AB* do not directly reveal any qualitative, intrinsic physical relations between *A* and *B*. But this does not imply that there are no dynamical relations between them. Some of these supervene on the qualitative, intrinsic physical properties of *A* and of *B*. Just as in classical particle mechanics, there may be other, nonsupervenient, concealed intrinsic dynamical relations, including spatial relations. While in classical particle mechanics these may be revealed by moving from a phase-space representation to a direct specification of the relative spatial positions of *A* and *B* (together with other of their dynamical properties such as momentum), on the interpretation of quantum mechanics presently under consideration, however, this is not so straightforward. On this interpretation, specifications of the dynamical states of *A* and *B* cannot be taken to include a spatial relation corresponding to a precise value for the distance between *A* and *B*.[24] On the other hand, they do include "weaker" spatial relations, such as "not being spatially coincident" or, perhaps, "being at least ten meters apart." Thus, while there are intrinsic spatial relations between *A* and *B*, these cannot be derived from a single fundamental relation of being separated by a certain definite distance.

---

[23] Including local, hidden variable theories and perhaps also Einstein's interpretation. Whether Einstein's interpretation is, in fact, ruled out on these grounds remains controversial.

Nonseparability follows immediately. For assume that *A* and *B* are ultimate, and so the only nontrivial, parts of *AB*, and that no other objects figure in some process involving the total spin of *AB*. Since *A* and *B* are not spatially coincident, this process cannot supervene on an ascription of qualitative, intrinsic physical properties at their common space-time locations. But if it is to supervene on any ascription of qualitative, intrinsic physical properties at space-time points, then these must be ascriptions of properties to *A*, *B*, or *AB*. Hence, no such process can be supervenient upon an ascription of qualitative, intrinsic physical properties at space-time points where that process is occurring.

Pure physical holism may be argued to hold as follows. On this interpretation, intrinsic physical properties of *AB* like total spin do not supervene even on the qualitative, intrinsic physical properties and relations of *A*, *B* corresponding to those which figure in the supervenience basis for the corresponding classical system, suitably enriched by the new qualitative, intrinsic physical properties of *A*, *B* associated with the nonclassical quantity intrinsic spin. Since the move from classical to quantum mechanics otherwise weakens the supervenience basis (for example, by removing a relation corresponding to precise relative positions of *A*, *B*), these properties of *AB* clearly also fail to supervene on the qualitative, intrinsic physical properties and relations of *A*, *B* postulated by quantum mechanics.

For example, *AB* has total spin zero. That *AB* has this property follows, on the present interpretation, from the fact that its state vector in the spin-singlet state lies in the subspace corresponding to this property (and, consequently, a measurement of total spin will, with probability one, reveal that it is zero). But *AB*'s possession of this property does not supervene on qualitative, intrinsic physical properties and relations of *A* and *B* separately. In particular, it does not supervene on the individual spins of *A* and of *B*, since (on this interpretation) *A* and *B* have no nontrivial intrinsic spin properties in the singlet state. *A*, *B* are basic parts of *AB*, relative to any class of processes that preserve systems of these types. Consequently, we have an instance of pure physical holism.

---

[24] A necessary condition for it to include such a relation is that the quantum-state vector of the joint system *AB* lie in the subspace corresponding to some precise value for the difference in coordinates of *A* and *B* along some direction. The wave function introduced in the famous paper by Einstein, Podolsky, and Rosen may seem to satisfy this condition. But, in fact, it does not, since the relevant subspace in this case actually fails to contain any nonzero vectors.

It might be objected against this argument that *AB*'s having spin zero is in fact supervenient on a certain qualitative, intrinsic physical relation *S* between *A* and *B* which might be described as "having equal but oppositely directed spins." Now this description of the supposed relation is unfortunate, since on the present interpretation neither *A* nor *B* has any definite spin in the singlet state. But perhaps quantum mechanics should be understood to postulate such an intrinsic physical relation, however it is referred to? If so, then no instance of pure physical holism has been presented, since *AB*'s having spin zero is determined by the fact that *A*, *B* bear one another this relation.

Notice that this objection is closely related to the second a priori argument against pure physical holism considered and rejected in section III. That the supposed relation *S* holds between two objects would not suffice for the pair *AB* to exist. But, given that the pair *AB* exists, this relation's holding between *A* and *B* would be necessary and sufficient for *AB* to possess the property *J* of "having spin zero."

But, in fact, there are reasons for denying that *S* is an intrinsic physical relation postulated by quantum mechanics. It is not an accident that, while "having spin zero" is an important quantum mechanical property of the pair *AB*, the relation *S* is not considered significant enough to warrant a name. The total spin of a system is a physical property that is fundamental to the behavior of that system in quantum mechanics. Its spin is determined by the spins of its component systems, together with a way of combining these spins (the vector addition law). But the property so determined is of the very same kind as the properties that help to determine it. As far as its spin goes, it is irrelevant whether or not a system is composed of subsystems: quantum mechanics applies to the spin of a system in just the same way in either case. This is important, since it permits one to treat the total spin of a complex system like a silver atom just as one would that of a spin ½ system with no nontrivial subsystems.

These considerations provide weighty reasons why total spin is an intrinsic physical property of a quantum system, whether that system is simple or compound. Of course, given that a pair of spin ½ systems has total spin zero, it follows that the systems in that pair are related by the physical relation *S*. But there is no reason to believe that *S* is an intrinsic rather than a derivative physical relation.[25] This answers the objection to the claim that such pairs establish pure physical holism on the interpretation of quantum mechanics presently under consideration.

It is worth pointing out that there are interpretations according to which quantum mechanics demonstrates neither nonseparability nor pure physical holism. These include interpretations of quantum mechanics involving nonlocal, hidden variables. Such interpretations have not been widely accepted. But here we have a clear parallel with the status of indeterminism in quantum mechanics. For a nonlocal, hidden variable interpretation of quantum mechanics might, if accepted, render quantum mechanics compatible also with determinism. Nevertheless, the impact of quantum mechanics has prompted contemporary metaphysics to accept the possibility of physical indeterminism. It must similarly come to terms with the possibility of physical holism and nonseparability. We could know for sure whether or not quantum mechanics demonstrates either nonseparability or pure physical holism only if we both understood and believed quantum mechanics. Metaphysics can and must proceed without this knowledge.

RICHARD A. HEALEY

University of Arizona

---

[25] Notice that *S* counts as an external relation on Lewis's classification (see *On the Plurality of Worlds*, p. 62). This suggests that he would classify *S* along with spatial relations as an intrinsic relation. But this points to a difficulty with his scheme. Lewis divides properties into intrinsic and extrinsic based on whether things have them in virtue of the way they themselves are, or in virtue of their relations or lack of relations to other things. An analogous division among relations would count spatial relations, but not *S*, as intrinsic. For *A* and *B* have *S* in virtue of the way they are related to *AB*, rather than in virtue of the way they are related to one another.

BOOK REVIEWS

Sources of the Self. CHARLES TAYLOR. Cambridge: Harvard University Press, 1989. xii, 601 p. Cloth $37.50.*

Modern moral philosophy, Charles Taylor tells us, has suppressed or "occluded" discussion of the human good. Since he finds "[t]he very claim not to be oriented by a notion of the good" to be "incredible" (489), Taylor sees this as a failure. Ignoring the work of G. E. Moore, Max Scheler, Nicolai Hartmann, R. B. Perry, and other value theorists, he offers a remarkably original approach to the topic. The self, on Taylor's view, is not an objectively given entity like a star or a bacillus. Its identity is constituted largely by its values. To study the good is thus to study what makes us the selves we are. And we cannot study the sources of the self statically: only a historical narrative will enable us to understand the variety of goods that provide the ultimate rationales for our present ways of life and thought.

Taylor's study of the good thus goes beyond mere value theory. It is shaped to continue his long-standing battle against reductivist and naturalist epistemologies and ontologies generally. As he has explained elsewhere, he thinks the scientific and epistemological arguments for such views are extremely weak. They really draw their appeal from unavowed moral grounds. Other critics, from Nietzsche through Michel Foucault, have tried to show this, but their own positions are "often underdemonstrated, indeed rather impressionistically argued for." Moreover, Taylor thinks that we cannot just dismiss the naturalistic view of the self. It is "far from being simply wrong" and we are all "deeply imbued with it." We must rather try to see its limits by an examination of "the spiritual roots of naturalism" which shows "how, through the whole course of the development of the modern identity, the moral motivation has been intertwined with the epistemological." The latter by itself would never have been sufficient to win the support that reductionism and naturalism now have. "What would ultimately carry conviction" for his view, Taylor claims, "would be an account of this development which illuminated it and made more sense of it than its rivals, and particularly than naturalistic ones. As for any hermeneutic explanation," he adds, "interpretive plausibility is the ultimate criterion."[1]

---

* My thanks to Edward Minar for helpful comments on a draft.

[1] *Human Agency and Language: Philosophical Papers,* Vol. I (New York: Cambridge, 1958), pp. 6–8. The whole "Introduction" and the first two papers in this volume form an almost indispensable prolegomenon to *Sources.*

0022-362X/91/8808/422–6

This helps us to see both the direction of Taylor's complex discussion in *Sources* and the nature of his argument. His narrative portrays the values of modernity evolving in interaction with its epistemologies and ontologies, each supporting and being supported by the other. He expects us to find that his history gives us a better understanding of the modern self than any explanation a reductionist can give. We shall then conclude that Taylor's account of the role of values in our lives is preferable to any account available to such theories. And this in turn will support Taylor's account of the domain of value.

No mean enterprise, this, and made more difficult by Taylor's own predilection for a theistic understanding of the ultimate good. St. Augustine, addressing God at the opening of the *Confessions,* said: "you made us for yourself and our hearts find no peace until they rest in you." Hobbes took our perpetual restlessness to be simply an indication of our being alive, not a result of mistaking earthly goods for that heavenly good that alone gives peace. Taylor is on St. Augustine's side. Although accepting many of the values of modernity, he thinks them in an important way inadequate: they impose demands without providing the motivational resources that would make compliance possible. All of us, apparently, would be dissatisfied if we would reflect on our situation. Taylor hints that only a return to a theistic moral source provides a way out.

In Part I, Taylor explains his idea of the self and of the "strong evaluations" that are necessary to it. Although our values are discovered through feelings, only a bad epistemology will suppose that this entails their entire subjectivity: our feelings reveal aspects of the world that are truly there for us (257). The "incomparably higher" goods thus made available to us enable us to assess our own desires and to find point and meaning in our ways of life and thought. We often are not aware of what these goods are, since articulating them requires work of a kind that reductionist philosophy discourages.[2] Presumably a satisfactory articulation is one that generates a hermeneutically plausible narrative. From here on, therefore, we get history.

Taylor discusses classical theorists (mainly Plato) and Augustine, but he is primarily concerned with developments from the end of the sixteenth century onward. The ancients are included to show that there was a time when sources of value were seen as external to the self. Augustine made a turn inward and since then values have been found increasingly by delving within the self, at first going through the self to find a source outside it (Taylor points to Descartes's proof of God's existence from the idea of God in his own mind) and ultimately finding the sources of value within the self as such. Two main lines of thought interest Taylor. Descartes and Locke moved toward seeing the self as a disengaged entity capable of controlling desire and of viewing everything other than itself as having at most instrumental value. The self is the autonomous chooser whose choices bestow value on everything else. The second line became prominent with Rousseau's rejection of this detached self and with the emergence of the Romantic expressive self. The self thus understood is first seen as the source of value when it is truly in contact with nature and then, more radically, as the source when it is expressing its own nature or constructing, through language, what it also expresses.

---

[2] 'Articulation' is a key term; for more explanation than is given in *Sources,* see "Self-Interpreting Animals" in the volume noted above.

What gives these versions of the self their appeal, Taylor holds, is that they seem to support or validate ways of living that have come to be especially valued in modern times. We believe in self-control and self-governance; we think increasingly of all people as equally deserving of respect; we value common life, not the elite lives of warrior heros, political leaders, or monks; we try to reduce suffering and pain, and so we demand of ourselves that we be benevolent always to everyone. If God was still a moral source for Descartes and Locke—if belief in God enabled them to see these values as motivating them to action—the tendency they represent toward finding sources within the self has now gone to extremes. The modern predicament is that the theories of the self we have accepted as supporting these values are coming to seem incapable of providing the moral sources of motivation we need in order to live as the values dictate. Taylor develops this thought largely through discussions not of philosophers but of post-Romantic literature.

For Taylor, then, the problem is not the one Alasdair MacIntyre described in *After Virtue*,[3] that we do not agree on common daily values. There are tensions between instrumentalists and expressivists, Taylor says, but all agree in placing high value on individual freedom, reduction of suffering, and universal benevolence and justice. This consensus makes our rare serious disagreements (for example, over abortion) stand out sharply (496). But our basic problem is that we cannot locate an incomparably higher good, through which we can come to take full satisfaction in the life goods to which we are committed (515ff). Such a good cannot be purely external: it must resonate within the subject, while not being merely subjective (510). It is presumably this need to find an incomparably higher good uniting us with an objective order which might be satisfied by a return to a theistic source; but Taylor barely suggests this (342, 521).

---

[3] Notre Dame: University Press, 1981.

Sources is a rich and rambling book, provocative and often illuminating as history, and philosophically interesting in showing us how Taylor's antinaturalist program works out in the domain of value. In the space available, I can only suggest a few questions.

First, Taylor talks as if the sole alternative to his view were a subjectivistic reductionism. But we might hold that, in the course of coping with practical problems, communities have developed multifaceted languages for use in a wide variety of contexts. If the languages of evaluation serve purposes that differ from those served by the languages of science, as Taylor agrees they do, then distinctions of subjective and objective can be sustained within the discourses of value, and need not be explained by some subsistent realm of the kind Taylor seems to suggest. On such a view, moreover, there need be no single best account of value. There would be numerous accounts whose goodness would depend on what one wanted the accounts to do.

Next, Taylor has difficulty making his historical account carry the philosophical weight he loads onto it. History is to show that constitutive goods disclosed through feeling have been the crucial factor in creating the distinctively modern identity. Even allowing that explanations invoking values may be illuminating although incomplete (and Taylor carefully points out their incompleteness), they must find some actual purchase in historical reality. Taylor, however, seems so certain that thought and action must ultimately be explained in terms of visions of the good that he does little to show either that major thinkers were themselves moved by values as he says they were, or that their ideas were widely accepted because of their ties with the values Taylor thinks they involved. For instance, he does not try to show that Locke's epistemology was seen by readers as being so weak that their acceptance of it has to be explained by something other than their belief in its strength as a theory of knowledge.

This is perhaps a small point; but Taylor's a priori assurance gives rise to a more serious defect in his narrative. He unaccountably neglects the question of why discussion of the *summum bonum* or of substantive goods was sidelined by many of the most influential modern philosophers in the first place. He provides no sustained treatment of modern natural-law theory, the work of writers from Grotius through Kant who tried to develop explanations of morality showing how people could live together despite deep disagreements about the good. Whether reductionists or not, they thought they could construct viable programs for social life precisely by ignoring the issues Taylor takes as central. The moral psychology to support this, which emerged most explicitly in Richard Price, Thomas Reid, and Kant, is that there is a previously unappreciated complexity to our motivation. We can be moved, according to the new view, not only by our grasp of the good but by our independent understanding of the right as well. This position, which responded to pressing practical problems, constitutes one of the major innovations of modern moral thought, so that the absence of any serious engagement with it leaves a chasm in Taylor's historical argument. Even if he is right, as I believe he is, in thinking that insight into practical principles helped to shape the epistemologies and ontologies of modernity, he has not shown that it was primarily principles about the good that had this force.

Third, from Taylor's Frankfurtian view about how human identity involves second-order desires capable of controlling first-order desires, nothing follows about any need for highest-order desires—for "ultimate" sources of "incomparably higher" value to give force to all the rest. Taylor presumably wants some such principle in order to move us from the values he and we share as modern selves to something we can only find in a theistic source of moral motivation. But he gives no account of what entitles something to be a source of value not needing any further backing. He thinks that, although the truth of the theistic position is often denied, "no one doubts that those who embrace it will find a fully adequate moral source in it" (317). Even if this astonishing remark were true (and it seems to ignore at least the many dimensions of the problem of evil), Taylor has done nothing to block the question of what incomparably higher value makes a theistic source adequate. Surely, if such a question can be raised about the adequacy of the values of human dignity, autonomy, and freedom from suffering, it can be raised about a theistic source, too.

Johns Hopkins University J. B. SCHNEEWIND

God, Time, and Knowledge. WILLIAM HASKER. Ithaca: Cornell University Press, 1989. xi, 209 p. Cloth $24.95. And *The Nature of God: An Inquiry into Divine Attributes.* EDWARD R. WIERENGA. Ithaca: Cornell University Press, 1989. ix, 238 p. Cloth $28.50.

William Hasker explores God's knowledge of the future. He is particularly interested in the relationship between such knowledge and human freedom. This is a clear book, filled with subtle and provocative argumentation. Hasker provides an argument against the possibility of "middle knowledge" and hence he rejects "Molinism." He offers an ingenious argument that "simple divine foreknowledge" has no point—it would be useless with respect to God's providential activity. Hasker also argues that various responses to incompatibilism about God's foreknowledge and human freedom (associated with Aquinas and Ockham) are inadequate. Finally, Hasker offers an alternative picture according to which the future is "open" in the sense that statements about future actions need not be true or false now. Although God knows the *possibilities* of the future and also the likelihoods that these will be realized, he cannot know about future free actions (insofar as there are no truths here to be known).

In this review, I cannot undertake to discuss various of Hasker's arguments. In particular, I shall not discuss his argument against the possibility of middle knowledge.[1] Also, I shall not discuss Hasker's claim that there is no coherent difference between "counterfactual power over the past" and "bringing about the past," and thus that a certain sort of compatibilism is incoherent.[2] I shall take issue instead with the following claims of Hasker: (1) the compatibilist must implicitly be operating with the notion of a "general power" rather than "the power to perform a particular act under given circumstances," and thus the compatibilist and incompatibilist are talking past each other; (2) the compatibilist cannot coherently affirm "lib-

---

[1] For a discussion of Molinism, see Alfred J. Freddoso, *On Divine Foreknowledge (Part IV of the Concordia)*, Luis de Molina, trans. (Ithaca: Cornell, 1988). Also, see the exchange between Hasker and Thomas P. Flint: Flint, "Hasker's *God, Time, and Knowledge*," *Philosophical Studies*, LX (1990): 103–16; and Hasker, "Response to Thomas Flint," *Philosophical Studies*, LX (1990): 117–26.

[2] I criticize this claim in "Introduction: God and Freedom," in my, ed. *God, Foreknowledge, and Freedom* (Stanford: University Press, 1989), pp. 18–23. For a discussion of this issue in the context of the argument for the incompatibility of causal determinism and freedom, see Carl Ginet, "In Defense of Incompatibilism," *Philosophical Studies*, XLIV (1983): 391–400; and my "Freedom and Miracles," *Nous*, XXII (1988): 235–52, esp. no. 5, p. 251.

0022-362X/91/8808/427–33

ertarian free will"—the claim that we sometimes have genuine alternative possibilities of choice and action.

I shall first lay out some of the considerations in virtue of which Hasker concludes that the compatibilist about God's foreknowledge and human freedom must be operating with the notion of a "general power." Note that one can have a generalized power while lacking the ability to exercise this power: I may be able (in the general sense) to play the piano but be temporarily paralyzed or tied down, and so forth. It would be highly damaging to compatibilism if it required this general notion of a power, since it is widely acknowledged that it is the notion of the power to perform a particular act in given circumstances (rather than the general power) that is relevant (in a straightforward way) to free will (and moral responsibility).

Hasker alludes to the distinction between hard and soft facts about times. He points out that there are two "moments or aspects" of the notion of a hard fact: first, a hard fact is "really about the past," and second, if a proposition expresses a hard fact, then it cannot be in anyone's power (at the relevant time) to render that proposition false (82).[3] Although Hasker describes these two "moments or aspects" of the notion of a hard fact, he *defines* hard facts as "true propositions, such that, with respect to a given time, it is impossible that anyone at or after that time should have the power to render them false." Given this definition of hard facts, Hasker explicates hard facts in terms of the notion of "future-indifference" and a number of other analytical hypotheses (84ff).

Having explicated the notion of a hard fact, Hasker argues that facts about God's prior beliefs about future actions are *hard* facts about the relevant past times (91–5). This turns out to be the case, on Hasker's approach, even though God's beliefs are not *temporally indifferent (nonrelational)* facts; since fixity and temporal indifference are different moments or aspects of hardness, a fact can be fixed even though it is not temporally indifferent.[4]

---

[3] I do not believe that "really about the past" captures the relevant temporal aspect, because a fact can be really about the past without being *solely* about the past: the conjunctive fact, "Jack sits at T_1 and Molly stands at T_3" is really (but not solely) about the past with respect to T_2. Thus, the notion of being really about the past does not capture the idea of being "fully accomplished and over-and-done-with," which is the appropriate notion.

[4] In a previous paper, Hasker distinguished two kinds of facts that correspond to the two moments or aspects of fixity and temporal nonrelationality: "Hard Facts and Theological Fatalism," *Nous*, XXII (1988): 419–36, repr. in Fischer, ed. This previous approach is, in my judgment, much preferable insofar as confusions are engendered by the "aspects" approach. For example, having defined hard facts as those which cannot be rendered false, Hasker says, "But even if the

Consider now Hasker's definition of the power to perform a particular act under given circumstances:

> In general, if it is in N's power at *T* to perform *A*, then there is nothing in the circumstances [fn. 18: It will be recalled that the circumstances that obtain at *T* comprise all and only the hard facts with respect to T] that obtain at *T* which *prevents* or *precludes* N's performing *A* at *T*. Here 'prevent' applies especially to circumstances that are *causally* incompatible with N's performing *A* at *T*, and 'preclude' to circumstances that are *logically* incompatible with *N*'s doing so (134).

On the basis of this account and his argument that God's beliefs are hard facts, Hasker concludes that the compatibilist must have in mind the general notion of power (and thus that the compatibilist and incompatibilist are talking past each other).

But Hasker's argument here is clearly unfair. He has produced an argument for the conclusion that God's beliefs are hard facts, but it is unfair to suppose that the compatibilist accepts this argument (or even knows of it). Indeed, the argument is controversial,[5] and its conclusion is precisely what is at issue in the debate between compatibilists and incompatibilists. Thus, the claim that God's beliefs are hard cannot simply be taken for granted when one is trying to see what notion of power the compatibilist is employing. Given this point, the compatibilist can embrace Hasker's account of the power to perform a particular act under given circumstances.

And there are other perfectly reasonable accounts of such power open to the compatibilist. *Any* theorist will want to adopt certain constraints on an adequate account of the "particular" notion of power. The compatibilist, it seems to me, can adopt the constraint that, in ascribing such power to perform *A*, one holds fixed all temporally nonrelational facts about the past and all present circumstances that prevent (in Hasker's sense) the agent's performing *A*. Thus, the compatibilist (who claims that facts about God's beliefs are temporally relational) can embrace a "particular" notion of power. Thus, Hasker has not given us reason to think that the com-

---

claim that God's beliefs are soft facts must be given up, it would be overly optimistic to expect the compatibilist at this point to fold his tents and steal silently away. For compatibilists have also made claims about powers we have over the past—claims that need not depend on the distinction between hard and soft facts . . ." (95). This remark makes sense if Hasker is talking about hard facts in the sense of the first moment or aspect—temporal indifference (or better, being "over-and-done-with," i.e., being solely about the past). But, given the official approach of the book, I do not see how it makes sense at all: if facts about God's beliefs are hard (on the current definition), tent folding seems in order for the compatibilist.

[5] See Hasker's comments on 92–3, esp. fn. 19.

patibilist is committed to an implausible conception of power or that the compatibilist and incompatibilist are talking past each other.[6]

Hasker claims that the compatibilist cannot consistently affirm what he calls "libertarian free will." (By this he means the sort of free will that requires genuinely open alternative possibilities; this might be contrasted with an "actual-sequence" approach to free will and moral responsibility.) This claim allegedly follows from an account of libertarian free will conjoined with Hasker's claim that the compatibilist is committed to the general notion of power.

Here is Hasker's account of libertarian free will:

> (FW′) N is free at *T* with respect to performing *A* = df. It is in N's power at *T* to perform *A* and it is possible at *T* for N to exercise that power, and it is in N's power at *T* to refrain from performing *A*, and it is also possible at *T* for N to exercise *that* power (138).

If one accepts (FW′) and also the claim that the compatibilist is committed to the general notion of power, then one must conclude that the compatibilist cannot affirm libertarian free will. This is because an agent's having the general power is compatible with its not being possible that he exercise it and, indeed, if God holds the relevant belief it is (allegedly) not possible for the agent to exercise the relevant power.

Hasker says:

> It is clear . . . that a compatibilist . . . is claiming that Peter can *have* the power to refrain from sinning *even though it is logically impossible that he should exercise that power under the existing circumstances*. But if one has the "power to do otherwise" only in that sense —the sense in which *having* the power does not guarantee that it is possible for the power to be *used*—then the central idea of libertarianism . . . has been lost (141).

But we saw above that the compatibilist need not embrace the general notion of power; rather, he can adopt a particularized notion. Also, the compatibilist can adopt a notion of possibility that permits the possibility of agents' exercising their powers, even if God exists. On the relevant notion of possibility, God's beliefs are not held fixed; but this is entirely appropriate (on the compatibilist's picture) insofar as God's beliefs are neither temporally nonrelational facts about the past nor present facts which are causally relevant to the agent's action.

---

[6] William P. Alston also believes that the compatibilist and incompatibilist are talking past each other; cf. "Divine Foreknowledge and Alternative Conceptions of Human Freedom," *International Journal for Philosophy of Religion*, XVIII (1985): 19–32; repr. in Fischer, ed. Alston's view is, however, slightly different from Hasker's. Alston does not attribute to the compatibilist the general notion of power; rather, he distinguishes between a compatibilist and incompatibilist "sense" of the particular notion of power.

Let us suppose that Peter sins at T. Is it possible for him to exercise his power to refrain from sinning? It is certainly impossible for him to exercise this power if *all* the "existing circumstances"—*including his sinning*—are held fixed. Thus, even an incompatibilist cannot hold such a stringent requirement. And when the requirement is relaxed (presumably employing the notion of soft facts), it seems that there is room for the compatibilist to assert that it is indeed possible for Peter to exercise this power. Thus, contrary to Hasker, there is no inconsistency in the compatibilist's affirming libertarian free will.

Wierenga systematically undertakes to give formulations of the various divine attributes, such as omnipotence, omniscience, perfect goodness, and so forth. He carefully sets out the accounts, and then evaluates them in the light of objections. In contrast to Hasker, Wierenga argues that God's omniscience is robust and includes foreknowledge of future free actions and middle knowledge.

In this review, I shall focus on Wierenga's discussion of the relationship between God's omniscience and human freedom. Whereas this discussion contains much that is of merit, I believe that it is vitiated by a problem similar to the one isolated above: Wierenga treats temporal nonrelationality (or being genuinely about the past) and fixity as different aspects of the generic notion of "accidental necessity." As with Hasker, this approach leads to certain confusions.

Here is a particularly glaring example of this sort of confusion. Wierenga criticizes a whole class of accounts of what he calls "accidental necessity," because they have as a consequence that certain facts such as "Socrates drank hemlock at T" turn out to be accidentally necessary. The problem, according to Wierenga, is that Alvin Plantinga has shown that there are possible circumstances in which such facts must not be considered accidentally necessary (99–100; 108).

I should point out, first, that the accounts in question (Freddoso, Widerker and Zemach, Fischer) all are putative accounts of temporal nonrelationality, not fixity *per se*. That is, they are all attempting to give accounts of hardness in the sense of temporal nonrelationality or being genuinely and solely about the relevant times. And facts such as "Socrates drank hemlock at T" are precisely this sort of fact. Now, whether all temporally nonrelational facts about the past are *fixed* is quite another matter, and Plantinga's examples pertain to this issue; thus, it is unfair to adduce these examples as part of a criticism of various accounts of temporal nonrelationality.

Further, let us consider Plantinga's examples more carefully. Plantinga[7] suggests that

> perhaps you will be confronted with a decision of great importance—so important that one of the alternatives is such that if you were to choose *it*, then the course of human history would have been quite different from what in fact it is. Furthermore, it is possible that if God had foreseen that you would choose *that* alternative, he would have acted very differently. Perhaps he would have created different persons; perhaps, indeed, he would not have created [Socrates]. So it is possible that there is an action such that it is within your power to perform it and such that if you were to perform it, then God would not have created [Socrates] (*ibid.*, p. 257).

Wierenga concludes that it is possible that a fact such as "Socrates drank hemlock at *T*" is not accidentally necessary (after *T*). Plantinga draws a similar conclusion. But even if we are considering the notion of "fixity" rather than temporal nonrelationality, this is dialectically unfair. For the dialectical context is one in which we are considering whether God's foreknowledge is compatible with human freedom. Thus, it is inappropriate to adduce in the manner of Plantinga and Wierenga an example in which it is alleged that some agent is free despite God's foreknowledge. Plantinga's examples, as far as I can see, make essential use of a foreknowing God (or at least some individual with relevantly similar attributes); but then they should not be invoked in a context in which it is controversial whether God's foreknowledge is compatible with human freedom. So, even if we restrict our focus to accidental necessity in the sense of fixity, the use of such examples is dialectically inappropriate.

Another worry. Wierenga points out that none of the accounts of accidental necessity he considers offers any reason to classify God's beliefs at a time as accidentally necessary. Thus, he concludes that the argument for incompatibilism based on the fixity of the past "must be judged to be a failure" (113). But I believe that Wierenga is too quick to leap to the compatibilistic conclusion. I believe that, even if God's beliefs at times are temporally relational facts about those times, they have a certain internal structure; further, this structure may include a temporally nonrelational or *hard* element. It may be that falsifying facts about God's beliefs requires affecting the hard internal component. Thus, there might be a perfectly good argument employing the fixity of the past to the conclusion that God's foreknowledge is incompatible with human freedom, even if God's beliefs are deemed soft facts.[8]

---

[7] Plantinga develops these examples in his "On Ockham's Way Out," *Faith and Philosophy*, III (1986) 235–69; repr. in Fischer, ed. Cf. Wierenga, 99.

Wierenga considers the response to the incompatibilist's argument that posits that God is atemporal. On this approach, His omniscience is not foreknowledge, and the argument from the fixity of the past cannot get off the ground. Wierenga argues against this atemporalist strategy that facts such as "God timelessly knows that Jones mows his lawn at *T*" obtain at times prior to *T* and can be considered hard facts about the times in question. (This point is also made by Plantinga (*op. cit.*, p. 239).)

But it seems to me that the facts in question are not hard facts about the relevant times. A hard fact about a time (in the sense of a temporally nonrelational fact about that time) is both *genuinely* and *solely* about that time. Some soft facts are genuinely but not solely about the times in question. Facts about God's timeless beliefs are not even genuinely about the relevant times: they do not allege that anything "really happens" at the relevant times. The distinction between hard and soft facts presupposes what Freddoso has called the "primacy of the pure present"; hard facts assert that something happens in the pure present, but facts about God's timeless beliefs make no such assertion. Thus, it is implausible to suppose that the Boethian strategy can be blocked in this fashion.

Finally, I remind the reader that, whereas Hasker argues that compatibilism (with a "particular" notion of "power") is incoherent, Wierenga defends it as true. For what it is worth, I disagree with them both: I believe that such compatibilism is coherent but false.

Both books are rigorous and contain many worthwhile and provocative arguments. They both apply sophisticated analytical tools to the subject matter. Despite the flaws noted above, I believe that they are enlightening and useful books. They constitute worthy additions to Cornell University Press's series, *Cornell Studies in the Philosophy of Religion*.

JOHN MARTIN FISCHER

University of California/Riverside

---

[8] I develop this sort of position in my "Hard-Type Soft Facts," *The Philosophical Review*, XCV (1986): 591–601.

Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. JUDEA PEARL. San Mateo: Morgan Kaufmann, 1988. 552 p. Cloth $49.95.

This is an extraordinary book. The author, an electrical engineer by trade, is writing for an audience of computer scientists concerned with artificial intelligence. Yet much of what he has to say is directly relevant to the concerns of philosophers. In fact, I would say that no philosopher concerned with probability, decision theory, defeasible logic, or even epistemology in its formal guises can afford to be ignorant of the contents of this book. Although it is written for researchers in artificial intelligence, and is formidably technical in some of its parts, it also contains much that is accessible to the nontechnical philosopher who possesses only a familiarity with elementary mathematics and logic.

The basic approach to uncertainty is a personalistic one. Probability is a function (a real-valued function) on statements. It is a subjective function, in the sense that there is no way in which an individual can "get it wrong." (Pearl does not discuss the problem of reconciling inconsistent personal probability assignments.) We can have the appearance of higher-order probabilities, because the proposition that an individual assigns a certain probability (has a certain belief in the occurrence of) an event is itself an event in the real world, and can have a probability.

This is not, nowadays, an unusual approach to probability, but Pearl seems to be quite unaware that this is hardly the "traditional" (358) approach. Similar jarring observations occur elsewhere. "Confidence intervals," for example, are used to refer to the intervals in which an individual's confidence lies, rather than to the common statistical construct of the same name. One should be aware of these things, but one should not complain: the adjustments to more conventional terminology are easy to make for those with the requisite background, which most philosophy students interested in these matters have.

The book covers a broad spectrum of issues, from a point of view, one component of which is the personalistic Bayesian view that a statement is probable because an agent thinks it so, and one component of which is to demand that only tractible computations are of serious interest. This latter point of view is an interesting one. Some decades ago what mattered in philosophy was what could be stated exactly and explicitly. That a certain sentence followed in a given logic from certain others was the paradigm: exact rules could be given. By the same token, that a certain sentence does *not* follow

0022-362X/91/8808/434–7

from certain others can be stated perfectly explicitly, even though there are no rules that tell us when this is the case.

We are being led, by artificial intelligence and computer science, to pay somewhat more attention to the processes by which we come to know that conditions such as "provability" are fulfilled. It is one thing to have a clear idea of what it is to be a sentential tautology. It is quite another matter, even though the question is perfectly decidable, to have a practical way of deciding whether an arbitrary formula, with, say, a few hundred components, is a tautology.

In the same way, it is one thing to adopt a subjectivistic or personalistic view of probability, and smilingly say: well, it is just a matter of computing your posterior; it is quite something else to explore ways of being able to do that in *reasonable* amounts of time. As a computer scientist, Pearl focuses on the latter. Whether philosophers should do likewise is an open question—one might most plausibly suppose that they ought to explore *all* questions—but it is certainly refreshing and important to be reminded of the difference between "in principle" and "in practice." An instance in which the distinction between in-principle definability and computational feasibility may be deeply important is the possible role of computational feasibility in determining how people actually think about the world. Pearl claims, for example, that "the human obsession with causation . . . is computationally motivated" (383). This idea is explored for several pages, and it certainly seems a fruitful one to consider, going far beyond the simple notion that people "invent" the causal relation. It may be able to explain how and when causal relations are interposed between observational events.

Pearl's book provides numerous techniques for getting on the right side of the line between tractable and intractable computation. One of the most important and effective is the technique of exploiting "conditional independence." This is a notion that should play a frequent and fundamental role in philosophy. Although the *idea* is familiar to philosophers—*A* and *B* are conditionally independent given *C*, if, once you know *C*, the fact that *A* occurred does not tell you anything new about *B*. For example, grain production and golf scores are not independent: high grain production tends to be associated with high golf scores; but if we know the weather, the association is explained away: given a rainy season, there will be a lot of grain produced, and there will be high golf scores, given a rainy season, quite independently of whether in addition there is a lot of grain produced.

It turns out that this idea of conditional independence plays an important role in many realms: defeasible reasoning, nonmonotonic logic, epistemic updating, knowledge representation, etc. It can also be exploited in developing decision trees. Many of these areas are explored in Pearl's rich book.

The first chapter of Pearl's book provides an introduction to the importance of probability in AI systems: something few philosophers will need to be persuaded of. It is not the existence of uncertainty that is important for philosophers, but the question of how to handle it, how it is related to classical axioms for probability. Chapters 2, 3, 4, and 5 concern the representation of belief and its computationally feasible revision. These chapters contain the technical meat of the book, as well as containing a serious look at issues that are not so comfortable for Pearl's approach. These chapters will perhaps be the most difficult going for philosophers. Part of the reason for this is the unfamiliarity of the terminology: some familiar philosophical ideas are given computer science names (278: ". . . more plausible, i.e. . . . has the highest posterior probability"). The same is true of the examples, for example, medical diagnosis, but here their unfamiliarity may provide the means for communicating new ideas.

Chapter 6, on decision and control, may well be one of the most rewarding in the book for philosophers. It is one thing to contemplate the abstract possibility of a complete decision tree for a moderately complex decision, but it is quite another matter to use these ideas in a responsible and tractable way to improve the quality of our decision making. There are no conceptual breakthroughs here —no challenges to the standard axioms. But the passage from abstract possibility to practical reality can have a profound effect on our view of decision theory. Here, as in other places, the one drawback of wordprocessing makes itself felt: inconsistency among the iterations of a manuscript. For example, a proof on page 296 refers to the "decomposability axiom," but there is no such axiom. One can figure out which axiom is intended among those a page or so earlier (axiom 3), but it is frustrating to have to do so. Such obstacles notwithstanding, the chapter is a brilliant one.

The seventh chapter will be seen as one of the weakest, from the point of view of the philosopher. Pearl has read widely and well in the philosophical literature as well as in the statistical literature. Nevertheless, there remains a certain incompleteness and inaccuracy that occasionally jars. The eighth ("Learning Structure from Data"), by contrast, provides material that is unfamiliar to most philosophers, and that should be familiar. It does not concern what we think of as induction, despite its title, but rather the problem of how best to structure a domain, given extensive statistical data about that domain. Take statistical inference for granted; suppose you have as much data as you want: now what can you learn about causal relations, about independence, and the like.

Chapter nine will give many philosophers a feeling of *deja vu*. Here again is Shafer's novel representation of uncertainty, this time dressed for artificial intelligence. As a thoroughgoing probabilist, Pearl must bend over backward to give the Shafer/Dempster formalism (as it is referred to in AI) its due. In fact, he bends over rather too far, and is more concerned to illustrate differences than to find commonalities. Thus he finds a "semantic" difference, in that Shafer's view is concerned with "degrees of provability," while probability is concerned with "degrees of belief." It is not clear why construing 'Bel' as the lower envelope of *certain* classes of probability functions does not resolve the differences; Pearl does not address this question.

The final chapter, on the relation between probability and various nonstandard logics: logics for defeasible reasoning, default logics, and the like, will strike many philosophers as the most fun and most informative. Many of the novelties of logic in artificial intelligence are revealed here; it is an area in which philosophers have much to contribute.

Finally, there is an extensive bibliography.

This is a great book. It is not always easy. Its difficulty is compounded by the fact that the writing and editing are not always smooth (a difficulty that may, by now, have been corrected in a new edition). It will nevertheless be a rewarding book for most philosophers, as well as statisticians, decision theorists, logicians, probabilists, and, of course, those in artificial intelligence.

HENRY E. KYBURG, JR.
University of Rochester

Knowledge in Flux: Modeling the Dynamics of Epistemic States. PETER GÄRDENFORS. Cambridge: MIT Press, 1988. xi, 262 p. Cloth $27.50.

Recently an increasing number of writers have sought to direct the attention of epistemologists to the topic of belief change. Thus, Peter Gärdenfors declares his intention in the preface of his book to examine "how to change your mind." The result of his efforts is an outstanding piece of work.

0022-362X/91/8808/437–44

Discussing belief change presupposes a conception of the space of changes available to an inquirer in a given state of belief and, as a consequence, a characterization of the space of belief states to which the inquirer can move. Gärdenfors's discussion of this matter in chapter 2 offers an overview of diverse "models" that have been proposed by philosophers and computer scientists in recent years.

Most of Gärdenfors's discussion focuses on what I would call changes in "states of full belief." What is fully believed is taken for granted by the inquirer as settled, at least for the time being. There is "no doubt" that its negation is false (23). Hence, the inquirer is maximally certain that it is true and assigns it subjective or credal probability 1. Gärdenfors takes for granted, however, that full belief is not only sufficient for assigning probability 1 but necessary as well, and develops an overview of Bayesian models of epistemic states in chapter 2, section 7, and, on this assumption, accounts of change in probability judgment in chapter 5.[1]

Chapter 3 investigates some of the structural properties of changes in states of full belief due to adding new information (perhaps through observation, testimony, or inference) or removing information already taken for granted. In expansions, information is added to the state of full belief together with all of the implications thereof, regardless of whether this issues in inconsistency or not. Nothing is given up. Gärdenfors thinks inconsistency is "epistemic hell" (51), however. When an agent adds some information to his initial epistemic state which is inconsistent with it, the inquirer should not retain everything in the initial state as he would in expansion but should give up the least amount of information needed to avoid contradiction while adding the new items. Thus, when an inquirer changes his epistemic state due to the addition of new information, Gärdenfors calls it a "minimal revision." Minimal revisions are expansions when the new information added is consistent with the initial state; but, otherwise, minimal revisions require the inquirer to give up some old information in order to ensure consistency.

Sometimes the initial "input" demands that a proposition be given up without doing so in order to avoid the epistemic hell of inconsistency. Such changes are "contractions." Just as the addition of inconsistent information induces secondarily the removal of information, the deletion of information induces secondary deletion. Revision and contraction both involve ceasing to believe what one initially believed due to the pressure of adding information inconsistent with the initial doctrine, in the case of revision, or due to the pressure of initially removing information, in the case of contraction. Such secondary deletions call for choosing which portions of the old information to remove, and this involves evaluating assumptions one initially takes for granted (and, hence, initially assigns probability 1) with respect to their vulnerability to being given up or, alternatively, how well they are "entrenched" in one's initial state of belief. Gärdenfors recognizes that entrenchment is not a question of assessing how probable a claim is and, in chapter 4, he investigates some of the structural properties that evaluations of entrenchment should have. Chapter 5, as already stated, explores changes in probability judgment.

---

[1] I doubt the necessity of full belief for a judgment of probability 1. A person could reasonably assign probability 0 to the hypothesis that a piece of meat weighs exactly 3 pounds (and, hence, probability 1 to its negation) while remaining in doubt as to whether the weight is exactly 3 pounds or not. Indeed, the person could fully believe that the weight falls in the interval from 2.5 lbs to 3.5 pounds while being in doubt as to what the exact value is. He could do so while assigning 0 credal probability to every exact "point value" in the interval. All of this stands in contradiction to the necessity of full belief for probability 1.

Part II of the book, consisting of four further chapters, considers various applications of the ideas developed in the first five chapters. Chapter 6 develops the idea that propositional logic can be rationalized by imposing certain plausible requirements on how expansions ought to proceed. Chapter 7 examines belief-change approaches to the analysis of conditionals along lines pioneered by F. P. Ramsey. Chapter 8 contains an account of an epistemological approach to explanation where changes in belief, especially contractions, play a central role. Chapter 9 attempts a similar approach for causal beliefs.

The discussions are clear, informative, and responsible throughout, even when the positions advanced are controversial. The result is the best book-length overview of the formal aspects of belief change available.

Gärdenfors's favorite way of representing states of full belief linguistically associates with each potential state a set of sentences in the (regimented) language L. Gärdenfors supposes that L contains the standard truth-functional connectives, that it is governed by a consequence relation $\vdash$ between a set of sentences in L and a sentence in L. If A is a truth-functional tautology, then $\vdash A$. If $\vdash A \leftrightarrow B$ and $\vdash A$, then $\vdash B$. It is not the case that $\vdash \bot$. $\vdash$ satisfies the deduction theorem and compactness. He then defines a nonabsurd belief set to be a set of sentences closed under the consequence relation for L and consistent according to that consequence relation. He takes for granted that legitimate changes of states of full belief are representable as changes in nonabsurd belief sets.

Notice that changes in belief sets are understood as representations of changes in belief states. For Gärdenfors, the object of study is not transformations of linguistic entities but of "epistemic states" (states of full belief). In chapter 6, Gärdenfors exploits this idea to sketch a characterization of propositions that avoids explicating them either in terms of linguistic entities or possible worlds but rather construes them to be functions from epistemic states to epistemic states. Gärdenfors begins with a belief model taken to be an ordered pair ⟨K, Prop⟩ consisting of a set K of epistemic states and a set Prop of functions from K to K. Given an epistemic state K in K and a proposition A in Prop, $A(K)$ is intended to characterize the epistemic state K_A^+ which is the "expansion" of K by adding the proposition A. Gärdenfors introduces a system of assumptions and definitions about the kinds of functions in Prop which have this effect and which introduce the standard propositional connectives. In particular, given his assumptions, we can define a proposition B to be a consequence of a proposition A, if and only if for every K in K, $B(A(K)) = A(K)$. The postulates Gärdenfors introduces for the expansion functions in Prop suffice to induce a partial ordering of the epistemic states in K, although Gärdenfors does not explicitly say so. If B is a consequence of A, then $B(K)$ is a consequence of $A(K)$. Hence, one could have begun by characterizing a belief model as a set K of epistemic states and a consequence relation defined on this set. The expansions would then consist of all transitions from epistemic states to other such states that are consequences of the initial state. If, for some pair of epistemic states K and K', there is a greatest lower bound to the epistemic states that have both states as consequences (so that their "meet" exists), then the expansion of K by adding K' is that greatest lower bound or meet. Consequently, if we postulate that the meets of pairs of epistemic states always do exist, we can regard the epistemic states themselves as the propositions and represent the expansion of K by adding K' as the meet of these two epistemic states. Gärdenfors's approach does not guarantee that the meets of pairs of epistemic states always exist (though it does not preclude this). K is, as he explicitly states, completely unstructured except insofar as structure is induced by the functions in Prop. His postulates for Prop guarantee that the ranges of the functions in Prop (which are sets of epistemic states) have an algebraic structure. Given the seven postulates Gärdenfors finds compelling, they have the structure of a "pseudo Boolean algebra" sufficient to yield an "algebraic semantics" for intuitionistic propositional logic and, with an extra assumption, they have the structure of a Boolean algebra that yields classical propositional logic.

Gärdenfors's approach allows for the possibility that an inquirer may be in an epistemic state K and have available another state K' to which he might move without having available the conceptual opportunity of moving to the meet of these two states. In my judgment, this seeming generosity is, in point of fact, excessively restrictive on the conceptual opportunities available to the inquirer. One should not impose roadblocks in the path of inquiry by ruling out a shift to the meet. If this is right, there is no advantage in proceeding as Gärdenfors does, as compared to imposing the requisite algebraic structure (of a pseudo Boolean algebra or a Boolean algebra as the case might be) directly on the epistemic states and letting the epistemic states themselves serve as the propositions or "contents."

This reservation ought not to detract from Gärdenfors's achievement. If the topic is change in states of full belief, what we require in the way of logical structure ought to be derivable from our characterization of the range of epistemic states and the kinds of shifts that are thereby rendered available. Whether propositions are identified with epistemic states as I suggest or with expansion functions as Gärdenfors proposes, one can jettison much of the technical apparatus (such as the framework of possible worlds) that has dominated contemporary power without loss of power. Moreover, it becomes clear that linguistic representations of epistemic states by belief sets are just that—linguistic representations. Just as points in phase-space are used to represent mechanical states in classical mechanics without presupposing that mechanical systems speak the language of phase-space, the use of belief sets to represent epistemic states implies nothing about the way inquirers use language to manifest their epistemic states.

Gärdenfors resolutely eschews any concern with the concept of truth in his account of belief change. His pragmatism thereby resembles more closely the attitude of John Dewey than that of C. S. Peirce. This perhaps accounts, at least in part, for his ambivalence between resting content with an intuitionistic logic or a classical logic. For those like myself who see avoidance of error as a desideratum in inquiry, it is crucial that epistemic states or states of full belief be endowed with truth values, where truth or freedom of error is preserved under the consequence relation and where, for every epistemic state, there is a unique "classical negation" that is false (true), if and only if the first state is true (false). This kind of consideration argues in favor of requiring at a minimum that the epistemic states have the structure of a Boolean algebra and that the logic be classical. By allowing structures to be pseudo Boolean algebras that are not Boolean, Gärdenfors is recognizing that an agent may have a conceptual repertoire that is limited by the absence of negations for at least some epistemic states. For someone who thinks that avoidance of error is a desideratum in inquiry, the lack of an option of shifting to an epistemic state that is the classical negation of the initial one is an important deprivation of an epistemic option and, in Peirce's terms, does place roadblocks in the path of inquiry.

This objection presupposes that avoidance of error is an objective of inquiry. Given Gärdenfors's Deweyite indifference to truth, the worry being registered may not carry much weight with him.

Gärdenfors's belief sets (the linguistic representations of epistemic states) correspond to what elsewhere I have called "potential corpora" and modulo fine points of technical detail; there is very little difference between them. There is one important point on which we seem to disagree. Gärdenfors seeks an account of belief change that avoids countenancing any change to an absurd or inconsistent belief set or corpus. As a consequence, Gärdenfors has a rather different view from mine concerning what are the basic types of belief change.

I have advocated the view that all legitimate changes of states of full belief (changes of corpus) should be decomposed into sequences of legitimate expansions and contractions. It is a crucial feature of my view that sometimes expansion into inconsistency is legitimate. Gärdenfors's account of belief change seeks to preclude changes of this kind. He does so by taking revisions as basic. A revision of K by adding A is an expansion of K by adding A when A is consistent with K. It is what I have called a "replacement" in which $\sim A$ appears in the final state and A appears in the initial state otherwise. I deny that replacements are basic. To be legitimate, they must be decomposed into sequences of legitimate contractions and expansions. Gärdenfors does not, and, indeed, cannot require this.

The problem cases concern situations where the agent initially fully believes A and makes observations that yield reports that $\sim A$ is true or receives reliable expert testimony to this effect. According to my account, the agent expands into inconsistency from which he then seeks to extricate himself by contraction. He may do so by (i) questioning the process whereby he came to be sure that A is false, (ii) questioning the background assumption that it is true, or (iii) doing both. According to Gärdenfors's account, one never should expand into inconsistency in such cases. One should engage in a "minimal revision" that requires replacing A by its negation. This process will mimic my account when the contraction from inconsistency is of type (ii). Hence, if on my view the only type of legitimate contraction from inconsistency were of type (ii), the difference between Gärdenfors and myself would be a mere formal one. But I think it false to presystematic practice that, when one stumbles on a conflict between observation and background, one always questions the background rather than the observation. Nor do I think it unwise in some contexts to be skeptical of observation rather than or as well as the background. The interesting methodological issue is to identify the considerations that decide in favor of (i), (ii), or (iii) in any given context. Gärdenfors's approach precludes our addressing this issue by settling on a specific answer for all occasions in advance. For this reason, I think it an advantage to countenance expansions into inconsistency as sometimes legitimate and to regard expansion and contraction as basic.

Gärdenfors thinks of his investigations as addressing the "dynamics" of epistemic states. I think the analogy with physics is misleading in two important respects. First, as Gärdenfors clearly and rightly emphasizes, the epistemic states or states of full belief are best compared to equilibrium states. By representing such states by belief sets of sentences closed under a consequence relation, we are not assuming that inquirers are or prescribing that they ought to be logically omniscient. But as Gärdenfors rightly sees, the changes in view that occur when an agent recognizes a logical consequence of his current convictions, and becomes for this reason certain of its truth, are quite different in ways relevant to a systematic account of belief change from the changes that occur when an agent changes the set of convictions in a way that modifies their logical consequences. There are authors who do not recognize the difference as relevant or important. But Gärdenfors does, and I agree. Consequently, it is possible to regard the first kind of change as a change from disequilibrium to equilibrium and to envisage the inquirer as committed, while in disequilibrium, to realizing a certain equilibrium state. The second kind of change would then be regarded as a change in commitment to an equilibrium state. This is not the place to defend the superiority of this way of viewing things (which is Gärdenfors's way as well as mine) to views such as Gilbert Harman's. But if one embraces it, the account of belief change Gärdenfors and I are interested in is a study of changes from one equilibrium state to another. The principles regulating such changes (whether prescriptive or descriptive) are not analogs of dynamical principles but of principles of comparative statics. The name does not matter. But it is misleading to think of the kind of study Gärdenfors is engaged in, worthwhile as it is, as the analog of an investigation in dynamics.

This brings me to what, in my judgment, is the chief lacuna in Gärdenfors's magisterial essay. Gärdenfors restricts attention to changes where an initial "input" stipulates that a certain kind of change is to take place. The change may involve adding a proposition or removing one. In the case of expansion, addition of a proposition uniquely determines the new belief state. Just take the contents of the initial belief set and the sentence representing the new proposition and form the closure under the consequence relation. Gärdenfors bypasses the critical epistemological question of specifying conditions under which the agent is justified in adding some new proposition *A* to his initial belief state. This is not the usual foundationalist requirement that what is in the initial state be justified but only that the change contemplated be justified. According to the view of Dewey (whose views in some respects Gärdenfors seems to share), such changes are to be justified by looking at the goals of the inquirer and showing that the changes are best adapted to promoting the goals of the options available. There is no elaboration of this idea anywhere in Gärdenfors's book. The absence of attention to the decisional aspects of belief change carries over, at least in part, to Gärdenfors's discussion of contraction. Gärdenfors takes as input the demand to give up some proposition without considering the reasons that might justify attempts to gratify such demands. Gärdenfors recognizes the need to face the question of choosing between contraction strategies that remove the given proposition *A*, and has interesting things to say in this connection about entrenchment. This aspect of his view ought to fit naturally into a decision theoretic framework. Unfortunately, Gärdenfors does not recognize this. I conjecture that this failure accounts for his tendency to ignore difficulties with the so-called "recovery postulate" on contractions and with the "Harper identity" for revisions.

Even so, Gärdenfors has succeeded in making important contributions to understanding how to represent epistemic states and to identifying the salient properties of expansion, contraction, and revision. Space does not permit discussion of his views on conditionals and explanation; but his chapters on these matters ought to be taken seriously by philosophers of science and philosophers of mind. Overall, Gärdenfors's essay is a major contribution of interest to epistemologists, philosophers of science, philosophers of mind, students of artificial intelligence, and psychologists. It ought to change our minds about how to change our minds.

ISAAC LEVI

Columbia University

NEW BOOKS: ANTHOLOGIES

ABOULAFIA, MITCHELL, ed.: *Philosophy, Social Theory, and the Thought of George Herbert Mead.* Albany: SUNY, 1991. xx, 319 p. Cloth $39.50, paper $12.95.

ADAMS, MARILYN MCCORD, and ROBERT MERRIHEW ADAMS, eds.: *The Problem of Evil.* New York: Oxford, 1990. viii, 230 p. Cloth $59.00, paper $15.95.

AL-HIBRI, AZIZAH Y., and MARGARET A. SIMONS, eds.: *Hypatia Reborn: Essays in Feminist Philosophy.* Bloomington: Indiana UP, 1990. xii, 349 p. Cloth $39.95, paper $18.95.

ANNAS, JULIA, ed.: *Oxford Studies in Ancient Philosophy: Volume VIII 1990.* New York: Oxford, 1990. ix, 301 p. Cloth $34.50.

BEDAU, HUGO ADAM, ed.: *Civil Disobedience in Focus.* New York: Routledge, 1991. vi, 217 p. Cloth $69.95, paper $17.95.

BENHABIB, SEYLA, and FRED DALLMAYR, eds.: *The Communicative Ethics Controversy.* Cambridge: MIT, 1990. viii, 378 p. Cloth $37.50, paper $16.95.

BENTHAM, JEREMY: *Securities Against Misrule and other Constitutional Writings for Tripoli and Greece.* PHILIP SCHOFIELD, ed. New York: Oxford, 1990. li, 326 p. Cloth $98.00.

BERNASCONI, ROBERT, and SIMON CRITCHLEY, eds.: *Re-Reading Levinas.* Bloomington: Indiana UP, 1991. xviii, 252 p. Cloth $39.95, paper $14.95.

CARD, CLAUDIA, ed.: *Feminist Ethics.* Lawrence: Kansas UP, 1991. viii, 300 p. Cloth $29.95, paper $14.95.

COVER, J. A., and MARK KULSTAD, eds.: *Central Themes in Early Modern Philosophy: Essays Presented to Jonathan Bennett.* Indianapolis: Hackett, 1990. xiii, 336 p. Cloth $32.50, paper $16.95.

DAHLSTROM, DANIEL O., ed.: *Philosophy and Art.* Washington: Catholic UP, 1991. xii, 266 p. Cloth $44.95.

FISCH, MENACHEM, and SIMON SCHAFFER, eds.: *William Whewell: A Composite Portrait.* New York: Oxford, 1991. xiv, 403 p. Cloth $98.00.

FLANAGAN, OWEN, and AMELIE OKSENBERG RORTY, eds.: *Identity, Character, and Morality: Essays in Moral Psychology.* Cambridge: MIT, 1990. vii, 487 p. Cloth $37.50.

GEORGE, ALEXANDER, ed.: *Reflections on Chomsky.* Cambridge: Blackwell, 1989. xii, 271 p. Cloth $51.95, paper $19.95.

GOICOECHEA, DAVID, JOHN LUIK, and TIM MADIGAN, eds.: *The Question of Humanism: Challenges and Possibilities.* Buffalo: Prometheus, 1991. 341 p. Cloth $29.95.

HUNT, G. M. K., ed.: *Philosophy and Politics.* New York: Cambridge, 1990. vi, 147 p. Paper $16.95.

JOHNSON, GALEN A., and MICHAEL B. SMITH, eds.: *Ontology and Alterity in Merleau-Ponty.* Evanston: Northwestern UP, 1990. xxxiv, 206 p. Cloth $39.95, paper $12.95.

KEYT, DAVID, and FRED D. MILLER, eds.: *A Companion to Aristotle's* Politics. Cambridge: Blackwell, 1991. xii, 407 p. Cloth $63.95., paper $26.95.

LAMBERT, KAREL, ed.: *Philosophical Applications of Free Logic*. New York: Oxford, 1991. x, 309 p. Cloth $49.95.

LEPORE, ERNEST, and ROBERT VAN GULICK, eds.: *John Searle and His Critics*. Cambridge: Blackwell, 1991. xiv, 405 p. Cloth $44.95.

MACLEAN, IAN, ALAN MONTEFIORE, AND PETER WINCH, eds.: *The Political Responsibility of Intellectuals*. New York: Cambridge, 1991. xviii, 311 p. Cloth $49.50, paper $15.95.

MACNIVEN, DON, ed.: *Moral Expertise: Studies in Practical and Professional Ethics*. New York: Routledge, 1990. xix, 231 p. Cloth $55.00.

MORTLEY, RAOUL, ed.: *French Philosophers in Conversation*. New York: Routledge, 1991. 108 p. Paper $12.95.

PAKALUK, MICHAEL, ed.: *Other Selves: Philosophers on Friendship*. Indianapolis: Hackett, 1991. xiv, 273 p. Cloth $29.50, paper $7.95.

PERKINS, ROBERT L., ed.: *The Corsair Affair: International Kierkegaard Commentary: Volume 13*. Macon, GA: Mercer UP, 1990. xxv, 193 p. Cloth $24.95.

PETERSON, GRETHE B., ed.: *The Tanner Lectures on Human Values, vol. XII*. Salt Lake City: Utah UP, 1991. vi, 390 p. Cloth $30.00.

PINSENT, DAVID HUME: *A Portrait of Wittgenstein as a Young Man: From the Diary of David Hume Pinsent 1912–1914*. G. H. VON WRIGHT, ed. Cambridge: Blackwell, 1990. xxi, 119 p. Cloth $29.95.

REISS, HANS, ed.: *Kant: Political Writings*. New York: Cambridge, 1991. xv, 309 p. Cloth $44.50, paper $14.95.

ROSENTHAL, DAVID, ed.: *The Nature of Mind*. New York: Oxford, 1991. x, 642 p. Cloth $49.95, paper $19.95.

ROUSSEAU, G. S., ed.: *The Languages of Psyche: Mind and Body in Enlightenment Thought*. Berkeley: California UP, 1991. xix, 480 p. Cloth $55.00, paper $16.95.

SAYERS, SEAN, and PETER OSBORNE, eds.: *Socialism, Feminism and Philosophy: A Radical Philosophy Reader*. New York: Routledge, 1990. ix, 306 p. Cloth $69.95, paper $16.95.

SILVERMAN, HUGH J., ed.: *Writing the Politics of Difference*. Albany: SUNY, 1991. xvi, 372 p. Cloth $59.50, paper $19.95.

SIMON, YVES R.: *Practical Knowledge*. Robert J. Mulvaney, ed. New York: Fordham, 1991. xiii, 163 p. Cloth $35.00, Paper $19.95.

VILLANUEVA, ENRIQUE, ed.: *Information, Semantics, and Epistemology*. Cambridge: Blackwell, 1990. viii, 246 p. Cloth $49.95.

NEW BOOKS: PAPERBACKS

BATAILLE, GEORGES: *The Accursed Share: Volume 1: Consumption*. ROBERT HURLEY, trans. New York: Zone, 1991. 197 p. $12.95.

BRICKHOUSE, THOMAS C., and NICHOLAS D. SMITH: *Socrates on Trial*. Princeton: University Press, 1990. xiv, 337 p. $14.95.

BRAUDE, STEPHEN E.: *The Limits of Influence: Psychokinesis and the Philosophy of Science.* New York: Routledge, 1991. xiv, 311 p. $15.95.

CARRUTHERS, PETER: *Introducing Persons: Theories and Arguments in the Philosophy of Mind.* London: Routledge, 1991. viii, 265 p. $16.95.

CRANSTON, MAURICE: *Jean-Jacques: The Early Life and Work of Jean-Jacques Rousseau 1712–1754.* Chicago: University Press, 1991. 382 p. $17.95.

GELLNER, ERNEST: *Plough, Sword, and Book: The Structure of Human History.* Chicago: University Press, 1990. 288 p. $14.95.

HEIDEGGER, MARTIN: *Nietzsche: Volumes I and II.* DAVID FARRELL KRELL, trans. and ed. *Nietzsche: Volumes III and IV.* FRANK A. CAPUZZI, DAVID FARRELL KRELL, and JOAN STAMBAUGH, trans., DAVID FARRELL KRELL, ed. San Francisco: Harper Collins, 1991. 589 p. $16.95 (each).

HICK, JOHN: *An Interpretation of Religion: Human Responses to the Transcendent.* New Haven: Yale, 1991. xv, 412 p. $18.95.

JEFFREY, RICHARD C.: *The Logic of Decision:* 2nd ed. Chicago: University Press, 1990. xiv, 231 p. $14.95.

KIERKEGAARD, SØREN: *For Self-Examination: Judge for Yourself!* HOWARD V. HONG and EDNA H. HONG, trans. and ed. Princeton: University Press, 1991. xviii, 296 p. $14.95.

KURTZ, PAUL: *The Transcendental Temptation: A Critique of Religion and the Paranormal.* Buffalo: Prometheus, 1991. xvi, 500 p. $16.95.

MACGREGOR, GEDDES: *Dictionary of Religion and Philosophy.* New York: Paragon, 1991. xi, 696 p. $19.95.

PRICE, A. W.: *Love and Friendship in Plato and Aristotle.* New York: Oxford, 1990. xiv, 264 p. $18.95.

RAPAPORT, HERMAN: *Heidegger and Derrida: Reflections on Time and Language.* Lincoln: Nebraska UP, 1991. ix, 293 p. $10.95.

RAZ, JOSEPH: *Practical Reason and Norms.* Princeton: University Press, 1990. 220 p. $12.95.

RORTY, AMELIE OKSENBERG: *Mind in Action: Essays in the Philosophy of Mind.* Boston: Beacon, 1988. x, 378 p. $16.95.

ROSEN, STANLEY: *The Ancients and the Moderns: Rethinking Modernity.* New Haven: Yale, 1989. x, 236 p. $12.95.

SHANKER, S. G., ed.: *Godel's Theorem in Focus.* New York: Routledge, 1990. ix, 261 p. $14.95.

NEW BOOKS: TRANSLATIONS

ADORNO, THEODOR W.: *Notes to Literature: Volume One.* SHIERRY WEBER NICHOLSEN, trans. New York: Columbia, 1991. xv, 284 p. Cloth $35.00.

BAKUNIN, MICHAEL: *Statism and Anarchy.* MARSHALL SHATZ, trans. and ed. New York: Cambridge, 1990. xlix, 243 p. Cloth $44.50, paper $14.95.

BLOCH, ERNST: *Heritage of Our Times.* NEVILLE and STEPHEN PLAICE, trans. Berkeley: California UP, 1991. xiv, 377 p. Cloth $40.00.

BOURDIEU, PIERRE: *The Political Ontology of Martin Heidegger.* PETER

COLLIER, trans. Stanford: University Press, 1991. viii, 138 p. Cloth $22.50.

BOUSSUET, JACQUES-BENIGNE: *Politics Drawn from the Very Words of Holy Scripture*. PATRICK RILEY, trans. and ed. New York: Cambridge, 1990. lxxix, 415 p. Cloth $59.50.

CICERO: *On Duties*. M. T. GRIFFIN and E. M. ATKINS, eds. and trans. New York: Cambridge, 1991. li, 189 p. Cloth $37.50, paper $9.95.

DAVOUST, EMMANUEL: *The Cosmic Water Hole*. BARBARA JACHOWICZ, trans. Cambridge: MIT, 1991. xv, 206 p. Cloth $19.95.

DELEUZE, GILLES: *Empiricism and Subjectivity: An Essay on Hume's Theory of Human Nature*. CONSTANTIN V. BOUNDAS, trans. New York: Columbia, 1991. xi, 163 p. Cloth $32.50.

DUVE, THIERRY DE: *Pictorial Nominalism: On Marcel Duchamp's Passage from Painting to the Readymade*. DANA POLAN, trans. Minneapolis: Minnesota UP, 1991. xxii, 222 p. Cloth $39.95, paper $15.95.

HONNETH, AXEL, and HANS JOAS, eds.: *Communicative Action: Essays on Jürgen Habermas's* The Theory of Communicative Action. JEREMY GAINES and DORIS L. JONES, trans. Cambridge: MIT, 1991. ix, 301 p. Cloth $39.95.

KELSEN, HANS: *General Theory of Norms*. MICHAEL HARTNEY, trans. New York: Oxford, 1991. lx, 465 p. Cloth $110.00.

JACQUES, FRANCIS: *Difference and Subjectivity: Dialogue and Personal Identity*. ANDREW ROTHWELL, trans. New Haven: Yale, 1991. xxix, 374 p. Cloth $40.00.

JOHN OF SALISBURY: *Policraticus: Of the Frivolities of Courtiers and the Footprints of Philosophers*. CARY J. NEDERMAN, trans. and ed. New York: Cambridge, 1990. xxix, 240 p. Cloth $44.50, paper $14.95.

KLOSSOWSKI, PIERRE: *Sade My Neighbor*. ALPHONSO LINGIS, trans. Evanston: Northwestern UP, 1991. xv, 144 p. Cloth $24.95, paper $9.95.

KOFMAN, SARAH: *Freud and Fiction*. SARAH WYKES, trans. Boston: Northeastern UP, 1991. ix, 196 p. Cloth $35.00.

KRISTEVA, JULIA: *Strangers to Ourselves*. LEON S. ROUDIEZ, trans. New York: Columbia, 1991. xi, 230 p. Cloth $29.00.

PHILOPONUS and SIMPLICIUS: *Place, Void, and Eternity: Philoponus: Corollaries on Place and Void: Simplicius: Against Philoponus on the Eternity of the World*. DAVID FURLEY and CHRISTIAN WILDBERG, trans. Ithaca: Cornell, 1991. 153 p. Cloth $44.50.

SARTRE, JEAN-PAUL: *Critique of Dialectical Reason: Volume Two*. QUENTIN HOARE, trans. New York: Verso, 1991. x, 467 p. Cloth $60.00.

URVOY, DOMINIQUE: *Ibn Rushd (Averroes)*. OLIVIA STEWART, trans. New York: Routledge, 1991. iv, 156 p. Paper $18.95.

THE JOURNAL OF PHILOSOPHY

VOLUME LXXXVIII, NO. 9, SEPTEMBER 1991

EMPIRICAL EQUIVALENCE AND UNDERDETERMINATION*

DURING this century, there emerged from the philosophical analysis of scientific theories two results invested with broad epistemological significance. By the 1920s, it was widely supposed that a perfectly general proof was available for the thesis that there are always empirically equivalent rivals to any successful theory. Secondly, by the 1940s and 1950s, it was thought that—in large part because of empirical equivalence—theory choice was radically underdetermined by any conceivable evidence. Whole theories of knowledge (e.g., W. V. Quine's[1]) have been constructed on the presumption that these results were sound; at the same time, fashionable recent repudiations of the epistemic project (e.g., Richard Rorty's) have been based on the assumption that these results are not only legitimate, but laden with broad implications for the theory of knowledge.

In this paper, we reject both the supposition of empirical equivalence and the inference from it to underdetermination. Not only is there no general guarantee of the possibility of empirically equiva-

---

* The authors wish to thank Bas van Fraassen, Alan Goldman, William Lycan, Jim Maffie, David Stump, and Mary Tiles for criticism.

[1] The holistic theses of Pierre Duhem and Quine are probably the major instigators of this line of thought, as it appears in contemporary epistemology and philosophy of science. Quine's doctrines of the inscrutability of reference and the indeterminacy of translation are forms of underdetermination that depend on holism. See, e.g., "Epistemology Naturalized," *Ontological Relativity and Other Essays* (New York: Columbia, 1969), pp. 80ff.

The antirealist arguments of Arthur Fine and van Fraassen depend, in turn, on the thesis of underdetermination. See van Fraassen, *The Scientific Image* (New York: Oxford, 1980), ch. 3; and Fine, "Unnatural Attitudes: Realist and Instrumentalist Attachments to Science," *Mind*, XCV (April 1986): 149–79.

In the nineteenth century, both J. S. Mill and W. Whewell treated the possibility of empirical equivalence as an obstacle to scientific knowledge, differing as to whether and how it could be overcome.

0022-362X/91/8809/449–72

lent rivals to a given theory, but empirical equivalence itself is a problematic notion without safe application. Moreover, the empirical equivalence of a group of rival theories, should it obtain, would not by itself establish that they are underdetermined by the evidence. One of a number of empirically equivalent theories may be uniquely preferable on evidentially probative grounds. Having argued for these conclusions in the first two sections, respectively, we shall propose, in section III, a diagnosis of the difficulty that has impeded their recognition, and extract an attendant, positive moral for the prospects of epistemology.

I. PROBLEMS WITH EMPIRICAL EQUIVALENCE

A. Inducements to skepticism. The idea that theories can be empirically equivalent, that in fact there are indefinitely many equivalent alternatives to any theory, has wreaked havoc throughout twentieth-century philosophy. It motivates many forms or relativism, both ontological and epistemological, by supplying apparently irremediable pluralisms of belief and practice. It animates epistemic skepticism by apparently underwriting the thesis of underdetermination. In general, the supposed ability to supply an empirically equivalent rival to any theory, however well supported or tested, has been assumed sufficient to undermine our confidence in that theory and to reduce our preference for it to a status epistemically weaker than warranted assent.

Specifically, this supposed ability is the cornerstone of arguments for the inscrutability of reference and the indeterminacy of translation, which together insulate the epistemic agent by challenging the objectivity of criticism on which an entire philosophical culture has depended. It has spawned prominent, contemporary versions of empiricism, including those of Quine, Bas van Fraassen, and J. D. Sneed, which belie the promise of science to deliver theoretical knowledge. It encourages conventionalism in geometry through Hans Reichenbach's invocation of universal forces. It questions the possibility of ordinary knowledge of other minds through the contrivance of the inverted spectrum. It blocks inductive generalization through the stratagem of fashioning artificial universals to vie with natural kinds, as in Nelson Goodman's "grue" paradox, reducing the status of apparent laws to mere entrenchment.

The linkage between empirical equivalence and epistemic skepticism has roots that go back well beyond such contemporary manifestations. Hume reduced causal judgments to psychological projections of habit by offering coincidental concomitance as the empirically equivalent alternative to natural necessity. Descartes worried that, were our impressions illusory or demonically fabricated, what we take to be evidence would come out the same, inferring that empirical beliefs are unwarranted while those possibilities are open. Berkeley similarly exploited empirical equivalence to justify doubts about an external world. The implicit assumption in classical skepticism, operative still, is that no experience epistemically grounds a belief if that experience is strictly compatible with an alternative belief. Whence the method of undermining belief by constructing alternatives.

B. An argument against empirical equivalence. We find the pervasiveness of this influence out of proportion to the conceptual credentials of the basic idea of empirical equivalence. By connecting three familiar and relatively uncontroversial theses, we can construct a simple argument to cast doubt on empirical equivalence in general, as a relation among scientific theories (and, by parity of reasoning, between any rival perspectives).

On the traditional view, theories are empirically equivalent just in case they have the same class of empirical, viz., observational, consequences.[2] A determination of empirical equivalence among theories therefore requires identifying their respective empirical consequence classes. As the empirical consequences of any statement are those of its logical consequences formulable in an observation language, these classes are (presumably proper) subsets of the logical consequence classes of theories. Central, therefore, to the standard notion of empirical equivalence are the notions of observational properties, the empirical consequences of a theory, and the logical consequences of a theory. We shall show that, when these concepts are properly understood, the doctrine of empirical equivalence loses all significance for epistemology.

Our three familiar theses are these:

Familiar thesis 1, the variability of the range of the observable (VRO):

> Any circumscription of the range of observable phenomena is relative to the state of scientific knowledge and the technological resources available for observation and detection.

---

[2] Empirical equivalence can also be formulated in semantic terms: empirically equivalent theories have the same class of empirical models. Although the point of the semantic approach is to achieve independence of theory from language, one still needs a criterion of empirical status or observability to formulate empirical equivalence. One also needs to circumscribe the class of models with which a theory is to be identified, and this will require some reference to the theory's axioms or basic assumptions. For the question whether or not some particular set of structures is a model for a theory is not answered definitionally in science. It is answered by attempting to *apply* the theory, by working out consequences. Collateral information is crucial to such applications. For these reasons, we believe that questions about empirical equivalence must involve all the elements we shall bring to bear, whether the notion is formulated in ours or in semantic terms.

In particular, entities or processes originally introduced by theory frequently achieve observable or "empirical" status as experimental methods and instruments of detection improve. Such variability applies to any viable distinction between observational and theoretical language.[3]

Familiar thesis 2, the need for auxiliaries in prediction (NAP):

> Theoretical hypotheses typically require supplementation by auxiliary or collateral information for the derivation of observable consequences.

While direct derivability of statements bearing evidentially on theory is not in principle precluded, auxiliaries are generally required for the derivation of epistemically significant results.[4]

Familiar thesis 3, the instability of auxiliary assumptions (IAA):

> Auxiliary information providing premises for the derivation of observational consequences from theory is unstable in two respects: it is defeasible and it is augmentable.

Auxiliary assumptions once sufficiently secure to be used as premises frequently come subsequently to be rejected, and new auxiliaries permitting the derivation of additional observational consequences frequently become available.

Our argument against empirical equivalence now proceeds as follows. As VRO makes clear, the decision to locate a logical consequence of a theory outside its empirical consequence class (on the grounds of the former's nonobservational status) is subject to change. That class may increase, coming to incorporate an ever greater proportion of the theory's total consequence class.[5] This result already shows that findings of empirical equivalence are not reliably projectable, since we cannot reliably anticipate which of a theory's now unobservable consequences may become observable. But the problems with empirical equivalence run deeper than the inconstancy of the boundary of the observable. For even if it were possible to circumscribe the range of the observable relative to a state of science, we shall see that it would still be impossible so to circumscribe the range of auxiliary information available for use in deriving observational consequences.

---

[3] In labelling VRO "relatively uncontroversial," we acknowledge that van Fraassen's empiricism rejects it. Van Fraassen claims that what is observable is determined by facts about human beings as organisms, not by the transitory state of knowledge of those facts—not by science or technology. This view does not affect the use we shall make of VRO, however, because *judgments* of empirical equivalence must depend on *judgments* of what is observable; they cannot invoke transcendent facts.

But further, we reject the implicit assumption that conditions of observability are fixed by physiology. Once it is decided what is to count as observing, physiology may determine what is observable. But physiology does not impose or delimit our concept of observation. We could possess the relevant physiological apparatus without possessing a concept of observation at all. The concept we do possess could perfectly well incorporate technological means of detection. In fact, the concept of observation has changed with science, and even to state that the (theory-independent) facts determine what is observable, van Fraassen must use a concept of observation that implicitly appeals to a state of science and technology.

[4] An appeal to NAP in criticism of empirical equivalence may appear ironic in view of the role of holistic theses in fostering belief in empirical equivalence and underdetermination. It is indeed ironic that the problems NAP, in concert with other theses, poses for empirical equivalence have escaped the notice of holists.

By NAP, a theory's empirical consequence class must be allowed to include statements deducible from the theory only with the help of auxiliaries. One can distinguish the broad from the narrow class of a theory's empirical consequences, where the narrow class contains only observational statements implied by the theory in isolation from other theories and hypotheses. But NAP shows that it is the broad class, containing as well statements deducible only if the theory is conjoined with such auxiliaries, that matters epistemologically. Regardless whether holists are right in contending that the narrow class is empty, it is a class of little epistemic moment. It is by the complement of the narrow with respect to the broad that theories are primarily tested, and a characterization of empirical equivalence limited to the narrow would have no such epistemological consequences as we are concerned to contest.

It follows by IAA that, apart from shifts in observational status, a theory's empirical consequence class may increase through augmentations to the theory's total consequence class. As new auxiliary information becomes available, new empirical consequences derived with its help are added. Of course, conditionals connecting the auxiliary statements newly used to the empirical statements newly derived were already present among the theory's logical consequences. But the detached empirical statements are not present until the auxiliaries on which their deducibility depends become available. So long as we include within a theory's empirical consequence class statements derivable from the theory only via auxiliaries, so long as we construe that class *broadly*—and we have argued that it must be so construed to reflect the realities of theory testing—the theory's logical consequence class will be augmentable in virtue of containing the empirical consequence class as a subset. The empirical consequence class can also diminish, again by IAA, as the rejection of needed auxiliaries discontinues the derivability of some of its members. Therefore, any determination of the empirical consequence class of a theory must be relativized to a particular state of science. We infer that empirical equivalence itself must be so relativized, and, accordingly, that any finding of empirical equivalence is both contextual and defeasible.

---

[5] Conceivably, it could also decrease through a shift in the status of a consequence from observational to nonobservational, although that is not the usual pattern in the sciences.

This contextuality shows that determinations of empirical equivalence are not a purely formal, a priori matter, but must defer, in part, to scientific practice. It undercuts any formalistic program to delimit the scope of scientific knowledge by reason of empirical equivalence, thereby defeating the epistemically otiose morals that empirical equivalence has been made to serve. The limitations on theoretical understanding that a defeasible empirical equivalence imposes need not be grievous. Nevertheless, we think there is still less to the notion of empirical equivalence than survives these concessions.

It has been widely supposed that one can, utilizing the resources of logic and semantics alone, "read off" the observable consequences of a theory. The mobility of the boundary of the observable has been regarded as an inconvenience, not as a fundamental challenge to the idea that, at least in principle, the consequences or content of a theory are unambiguously identifiable. On this view, enhancements to our observational repertoire do nothing to alter a theory's semantics; rather, they merely shift the line, within the class of a theory's logical consequences, between observational and nonobservable consequences. But NAP shows that there is an epistemic question here quite distinct from the logico-semantic one. Specifically, the availability of auxiliaries—auxiliaries crucial for determining what a theory's empirical consequences are—is neither a matter of logic nor semantics; it is inescapably epistemic. The determination that a given empirical statement, *e*, is an empirical consequence of a particular theory, T, depends on whether there are epistemically well-grounded collateral hypotheses that establish a suitable inferential link between T and *e*. The availability of such hypotheses is clearly a matter of evidential warrant. Once statements whose derivability requires auxiliaries are allowed to count as consequences, as they must by NAP, no statement can be disallowed as a (broad) consequence of a theory unless some statements are disallowed as auxiliaries. So, before deciding whether a derivation of an observation statement from a theory plus auxiliaries qualifies the statement for inclusion in the theory's empirical consequence class, we must assess the epistemic standing of the auxiliaries.

This makes clear that the epistemic bearing of evidence on theory is not a purely logical relation, but is subject to reinterpretation as science grows and may be indeterminate at a particular point in the process of growth. How well supported an auxiliary is by evidence available *now* may depend on findings made *later*—a problem exacerbated by the fact that standards of evidential support themselves are transformable by the fortunes of empirical beliefs.[6] This fact darkens the prospects even for time-indexed delineations of empirical consequence classes.

C. *Response to anticipated objections*. The response we anticipate to our argument is a challenge to its assumption that empirical consequence classes must be identified for their equivalence to be established. Can there not be a general argument to show that classes must be the same independently of determining their membership? An obvious suggestion is that logically or conceptually equivalent theories must have the same consequence class, whatever that class is. As we do not question the empirical equivalence of logically equivalent theories, we ignore this suggestion and assume henceforth that theories whose empirical equivalence is at issue are logically and conceptually distinct.

One approach to constructing a general argument is to invoke the Lowenheim-Skolem Theorem. This theorem asserts that any first-order, formal theory that has a model at all has a denumerable model. A standard proof uses terms involving individual constants indexed by the natural numbers as the domain of a model. But if the domain need only be a set of terms, it could just as well be any denumerable set whose members are proposed as the referents of those terms. So, in principle, such a theory has an infinite number of models.

The qualification 'in principle' must be emphasized here, because there is no guarantee that the denumerable models of a consistent, formal theory are effectively constructable. The proof of the theorem relies on the completability of a consistent first-order theory, and the complete theory need not be axiomatic even if the theory it completes is axiomatic. Thus, any appeal to the Lowenheim-Skolem Theorem on behalf of empirical equivalence shows at most that equivalent theories exist in principle; it does not show that they are entertainable as alternatives.

But there is a more fundamental objection to the relevance of the theorem. Having multiple models of a formal theory does not mean having multiple theories of common empirical content. A physical theory, by virtue of being a physical theory, includes a semantic interpretation of its formal structure; it is not simply a formal structure variously interpretable. A physical theory is inherently at least purportedly referential. Its referents may turn out not to exist, if nature fails to cooperate. But what its referents are if it has them is fixed by the theory itself; it is not a matter of optional interpretation. The reference-fixing devices of physical theory are rich and various. It does not require a philosophical theory of reference fixing to make the point that the semantic resources of physical theories transcend syntactic requirements on first-order formal theories. If, given the Lowenheim-Skolem Theorem, formal statements in first-order logic are referentially indeterminate, then a physical theory is not simply a set of formal statements in first-order logic.

---

[6] See our contributions to the symposium, "Normative Versions of Naturalized Epistemology," *Philosophy of Science*, LVII, 1 (1990): 20–34; 44–60.

Another approach is to construct an algorithm for generating empirical equivalents to a given physical theory, such as the Lowenheim-Skolem Theorem fails to do for formal theories. For example, there exist instrumentalist algorithms for excising the theoretical terms of a theory without empirical loss. Whether such algorithms are in fact successful is rendered highly dubious by the premises of our argument. It is by no means clear that a theory's instrumentalized version can match its capacity for empirical commitment, once the role of auxiliaries in fixing such commitment and the variability of the range of the observable are acknowledged. At most a theory's instrumentalized version can be held empirically equivalent to it relative to a circumscription of the observable and a presumed or intended domain of application. But while theories fix their own intended interpretations, they do not fix their own domains of application, nor the resources for detection of entities they posit. Algorithmically excised references may pick out entities that become detectable. New applications may arise with changes in collateral knowledge. Indeed, it is a measure of a theory's success when posited entities acquire a technological role, and applications for which the theory was not designed become possible.[7]

Be that as it may, what application of an instrumentalist algorithm to a theory produces is manifestly not an alternative *theory*. That is, the algorithm does not produce a rival representation of the world from which the same empirical phenomena may be explained and predicted. On the contrary, a theory's instrumentalized version posits nothing not posited by the theory, and its explanations, if any, of empirical phenomena deducible from it are wholly parasitic on the theory's own explanations. A theory's instrumentalized version cannot be a rival to it, because it is a logical consequence of the theory and is bound to be endorsed by anyone endorsing the theory. The challenge the instrumentalist poses is to justify endorsing more than the instrumentalized version, not to justify endorsing something instead of it. The point of the instrumentalist move is not to challenge the epistemic status of theory by demonstrating underdetermination among alternatives, but to argue that empirical evidence underdetermines theories individually by failing to discriminate between them and their instrumentalized versions. We know of no algorithm for generating genuine theoretical competitors to a given theory.[8]

---

[7] Consider the stature gained by Newtonian mechanics through its unforeseen applicability to the motions of fluids, and unforeseeable applicability to electric current.

The only other approach we know of to establishing empirical equivalence without identifying empirical content is to argue from cases. We propose an example, inspired by van Fraassen's in *The Scientific Image,* as representative. Let TN be Newtonian theory. Let R be the hypothesis that the center of mass of the expanding universe is at rest in absolute space. Let V be the hypothesis that the center of mass of the universe has constant absolute velocity v. Consider the claim that TN + R is empirically equivalent to TN + V.

This claim is based on the common TN component of the theories. It is Newtonian theory itself that assures us that unaccelerated absolute motion has no empirical consequences of a kind encompassed by the theory; that is, no consequences within mechanics. We can therefore bring two lines of criticism against the claim of empirical equivalence: either there is some other kind of consequence not envisioned within mechanics, or the underlying Newtonian assurance is wrong. The question is whether conceivable developments in scientific knowledge enable us to distinguish the theories empirically on one of these bases.

Van Fraassen's defense of his claim of empirical equivalence does not take account of these possibilities. He considers only extensions of the theories to further, late-breaking mechanical phenomena, arguing that whatever these may be, further equivalent extensions are constructable (*op. cit.*, p. 47ff). Can a defense be provided to cover the possibilities that van Fraassen neglects?

---

[8] We discount here the trivial algorithm that, applied to any theory T committed to theoretical entities of type r, generates the "rival" T* that asserts the world to be observationally exactly as if T were true but denies the existence of rs. Its logical incompatibility with T is insufficient to qualify T* as a genuine rival, as T* offers no competing explanations and is totally parasitic on T for whatever virtues it does offer. For extended criticism of such devices, see Leplin, "Surrealism," *Mind,* XCVI (October 1987): 519–24.

Imagine that TN + V has nonmechanical consequences absent from TN + R. A new particle, "the bason," is independently hypothesized to arise with absolute motion. The positive absolute velocity of the universe represents energy available for bason creation, and basons will appear under certain conditions in principle realizable in the laboratory if TN + V is correct, but not if TN + R is correct. v can be measured by counting basons, so that variants of TN + V for different v are empirically distinguishable.

We can construct an extension of TN + V which agrees with TN +R in not predicting basons. Let TN + W be TN + V plus the hypothesis that there is a velocity w such that basons appear if and only if and to the extent that $v > w$. Then the absence of basons establishes only that v does not exceed w; it does not require R. The presence of basons still refutes TN + R, but TN + R can be supplemented to allow basons; perhaps they arise spontaneously. TN + R then lacks an explanation of bason production, such as TN + V provides. Something in the way of explanatory parity is achievable by adding to TN + R the hypothesis that what absolute motion produces is antibasons, which immediately annihilate basons. So the presence of basons is explained by the lack of absolute velocity. Still, TN + R does not explain the frequency of bason detection, as TN + V does. The observed frequency must simply be posited, as a constant determined by experiment, and this procedure is an admitted disadvantage relative to TN + V. But this comparison does not affect empirical equivalence.

The appeal to nonmechanical, differentiating phenomena can be defeated, because, if empirical equivalence holds within mechanics, it continues to hold for any extensions of mechanics in which the presence or absence of additional, nonmechanical phenomena is made to depend on the value of a mechanical property. This seems to be a general result. If theories T_1 and T_2 are equivalent with respect to properties $p_1, \ldots, {}_n$, they have equivalent extensions for any enlarged class of properties $p_1, \ldots, {}_n, q_1, \ldots, {}_m$; where properties $q_1, \ldots, {}_m$ are functions of $p_1, \ldots, {}_n$. On the other hand, if $q_1, \ldots, {}_m$ are not functions of $p_1, \ldots, {}_n$, they cannot be used to discriminate between T_1 and T_2.

It remains to consider the common core TN of the allegedly equivalent theories of the example. Here our argument finds its proper vindication. The point to make is that this core, and with it the original rationale for regarding TN + R and TN + V as equivalent in the first place (that is, unextended to nonmechanical properties) is defeasible. We may have good or sufficient reason to regard theories as empirically equivalent, but there is no guarantee. That concession is all our argument requires. We do not deny the possibility that the world is such that equally viable, incompatible theories of it are possible. We do not deny the possibility of the world's being unamenable to epistemic investigation and adjudication, beyond a certain level. But whether or not the world is like that is itself an empirical question open to investigation. The answer cannot be preordained by a transcendent, epistemic skepticism.

It is noteworthy that contrived examples alleging empirical equivalence always invoke the relativity of motion; it is the impossibility of distinguishing apparent from absolute motion to which they owe their plausibility. This is also the problem in the pre-eminent historical examples, the competition between Ptolemy and Copernicus, which created the idea of empirical equivalence in the first place, and that between Einstein and H. A. Lorentz.[9] Either the relativity of motion is a physical discovery, founded on such evidence as the empirical success of Newtonian or relativity theory, or it is guaranteed conceptually. Both views have been advanced historically. Our reply to alleged examples of empirical equivalence takes the former view. We maintain not that there are no cases of empirical equivalence, but that the claim that there are is defeasible.[10] If the latter view is to be assumed, then we return to the caveat that we are not concerned to contest the empirical equivalence of conceptually equivalent theories.

II. UNDERDETERMINATION

We have argued that the thesis that every empirically successful theory has empirically equivalent counterparts is precarious, at best. But for now let us suspend our incredulity about empirical equivalence and suppose that the thesis is sound. We wish to explore in this section what, if anything, then follows from the existence of empirically equivalent theories for general epistemology.

A number of deep epistemic implications, roughly collectable under the notion of "underdetermination," have been alleged for empirical equivalence. For instance, it is typical of recent empiricism to hold that evidence bearing on a theory, however broad and supportive, is impotent to single out that theory for acceptance, because of the availability or possibility of equally supported rivals. Instrumentalists argue that the existence of theoretically noncommittal equivalents for theories positing unobservable entities establishes the epistemic impropriety of deep-structure theorizing, and with it the failure of scientific realism. Some pragmatists infer that only nonepistemic dimensions of appraisal are applicable to theories, and that, accordingly, theory endorsement is not exclusive nor, necessarily, even preferential. One may pick and choose freely among theories whatever works for the problems at hand, so that the distinction between theories and models is lost. In a phrase, the thesis of underdetermination, denying the possibility of adequate evidential warrant for any theory, has become the epistemic corollary to the presumptively semantic thesis of empirical equivalence.

---

[9] Contemporary examples are also limited to space-time theories, raising the possibility of underdetermination for certain topological features of space-time. For example, it might be possible to obtain the same consequences from a dense but discreet space-time as are obtained by adding dimensions to continuous space-time. A more general treatment of relative motion would subsume it under topological considerations.

[10] And we deny the omnibus a priori claim that every theory has empirically equivalent rivals.

Against these positions, we shall argue that underdetermination does not in general obtain, not even under conditions of empirical equivalence. As we have seen, empirical equivalence is chiefly seen as a thesis about the *semantics* of theories; underdetermination, by contrast, is a thesis about the *epistemology* of theories. It has been supposed that, if theories possess the same empirical consequences, then they will inevitably be equally well (or ill) supported by those instances. We shall contest this supposition and, with it, the reduction of evidential relations to semantic relations, on which it rests. We dispute the ability of semantic considerations to resolve epistemic issues. But even allowing the epistemic dimension we have discerned in empirical equivalence, we shall find that the relative degree of evidential support for theories is not fixed by their empirical equivalence.

Specifically, we shall show, first, that significant evidential support may be provided a theory by results that are not empirical consequences of the theory;[11] secondly, we shall show that (even) true empirical consequences need lend no evidential support to a theory. By this dual strategy, we propose to sever the presumed link between supporting instances and empirical consequences.[12] We shall show that being an empirical consequence of a hypothesis is neither necessary nor sufficient for being evidentially relevant to a hypothesis. These conclusions will establish that theories identical as to empirical consequences may be differentially supported, such that one is epistemically preferable to the other.

---

[11] Here we allow, as always, a role for auxiliaries in generating consequences.

[12] In "Realism, Underdetermination, and a Causal Theory of Evidence," *Nous*, VII (March 1973) 1–12. Richard Boyd sought to drive a wedge between empirical consequences and supporting instances. But Boyd's case rests on the (in our view dubious) principle that "new theories should, prima facie, resemble current theories with respect to their accounts of causal relations among theoretical entities" (p. 8). Our argument will require no such restriction.

A. *Evidential results that are not consequences*. We begin by noting that instances of a generalization may evidentially support one another, although they are not consequences of one another. Previous sightings of black crows support the hypothesis that the next crow to be sighted will be black, although that hypothesis implies nothing about other crows. Supposing this evidential connection to be uncontroversial, we ask why, then, in the case of universal statements it should be supposed that evidential support is limited to logical consequences.

Is it that the evidential connection admitted to hold among singular statements is at best indirect, that it connects those statements only via a general statement that they instantiate? The thesis would then be that *direct* evidential support for a statement is limited to its logical consequences, and singular statements instantiating the same generalization support one another only in virtue of directly supporting that generalization. In short, where there appears to be evidential support for a statement, *s*, outside the range of *s*'s logical consequences, such support is parasitic on support of a general statement, *m*, which entails *s*, from *m*'s logical consequences.

We believe this to be an unperspicacious way of accounting for what goes on in singular inference. Often the evidential link between singular statements is stronger than the support available for a general intermediary, whose identification can, in any case, prove elusive. But even if this account worked, it should be noted straightaway that allowing a statement to accrue indirect empirical support in this fashion already undermines the claim that statements are confirmable only by their empirical consequences. This result alone suffices to establish that the class of empirical consequences of a statement and the class of its prospective confirming instances are distinct.

We began this discussion with the hackneyed case of black crows in order to show that the possibility of inferences of even the most mundane sort (from particular-to-particular) depend upon denying the thesis that evidential support accrues to a statement only via its positive instances. This claim becomes even clearer when one considers the manner in which real scientific theories garner empirical support. Consider, for instance, the theory of continental drift. It holds that every region of the earth's surface has occupied both latitudes and longitudes significantly different from those it now occupies. It is thereby committed to two general hypotheses:

H_1: There has been significant climatic variation throughout the earth, the current climate of all regions differing from their climates in former times.

H_2: The current alignment with the earth's magnetic pole of the magnetism of iron-bearing rock in any given region of the earth differs significantly from the alignment of the region's magnetic rocks from earlier periods.

During the 1950s and 1960s, impressive evidence from studies of remnant magnetism accumulated for H_2. Clearly, those data support H_1 as well, despite the fact that they are not consequences of H_1. Rather, by supporting H_2 they confirm the general drift theory, and thereby its consequence H_1.

Similar examples are readily adduced. Brownian motion supported the atomic theory—indeed, it was generally taken to demonstrate the existence of atoms—by being shown to support statistical mechanics. But, of course, Brownian motion is no consequence of atomic theory. The increase of mass with velocity, when achieved technologically in the 1920s, supported kinematic laws of relativity which to that point continued to be regarded with great suspicion. J. J. Thompson's cathode ray experiments in the 1890s were important evidence for a host of theoretical hypotheses about electricity that depended on Lorentz's electroatomism. Phenomena of heat radiation were used by J. C. Maxwell in the 1870s to support the kinetic molecular theory, which did not address the transmission of heat energy across the space intervening between bodies. The emergence in the 1920s of evidence showing heritable variation supported Darwin's hypothesis about the antiquity of the earth, although that hypothesis entailed nothing about biological variation. Contemporary observational astronomy is replete with indirect methods of calculating stellar distances, whereby general hypotheses in cosmology acquire support from facts they do not imply about internal compositions of stars.

A number of points are to be noted about such examples. First, by dating them we emphasize that they are not dismissable by invoking auxiliaries via which the evidence is derivable. One could not in the 1890s represent Thompson's results as consequences of electrical laws by making electroatomism an auxiliary. Despite Ludwig Boltzmann's pioneering work, statistical mechanics was too speculative in 1905 to qualify as an available auxiliary. Even taking an ahistorical view, it would be casuistical to represent evidence as a consequence of a hypothesis from which it is derivable via auxiliaries, if it is the auxiliaries rather than the hypothesis that really fuel the derivation. If a formal criterion is wanted, we may stipulate that a hypothesis be ineliminable from the derivation of what are to qualify as its consequences.[13]

Second, the more general theory via which the evidence supports a hypothesis of which it is not a consequence need not be very precise or specific. For example, the statistical mechanics that Brownian motion supported was more a program for interpreting phenomenological thermodynamics probabilistically than a developed theory. There can be good reason to believe that conceptually dissimilar hypotheses are related such that evidence for one supports the other, without possessing a well worked out or independently viable theory that connects them. Perhaps a theory that connected them has been discredited without the connection it effected being discredited. In this respect, nonconsequential evidence for general statements approximates the case of singular statements for which the inferential link proved elusive.

Third, we need not fear running afoul of familiar paradoxes of confirmation in taking evidence to confirm a hypothesis in virtue of supporting a more general statement that implies the hypothesis. The intuition that what increases our confidence in a statement thereby increases our confidence in what that statement entails is fundamentally sound. The difficulties that Carl Hempel, for example, extracted from his "special consequence condition" depend on a certain logical form for general laws and a simplistic criterion of confirmation—Nicod's criterion—which requires, in opposition to the position we have undertaken to defend (see section II.B), that all positive consequences be confirming. Much sophisticated reasoning in the natural sciences would be vitiated by restricting evidence relevant in assessing a theory to the entailments (via auxiliaries) of the theory. And any singular prediction would be so vitiated as well.

Finally, we need to acknowledge and take into account a subtlety of confirmation that might appear to challenge the force of nonconsequential evidence for our argument. There is an obvious way in which a statement not entailed by a theory can be evidence for the theory. The statement might imply another empirical statement that *is* entailed. Suppose, for example, that the theory entails a—perhaps indefinitely extendable—disjunction of which the statement is a disjunct. By implying a statement that is a consequence, the evidence, though not itself a consequence, fails to discriminate between the theory and any empirically equivalent theory. So showing that there can be evidence for a theory that is not a consequence of the theory does not suffice to show that empirically equivalent theories can be differentially supported.

---

[13] Logico-semantic trickery will not render it ineliminable so long as epistemic conditions are imposed on auxiliaries.

Our examples, however, do not pose this problem. An example that does is to cite ten successive heads in tosses of a coin as evidence that the coin is biased. The hypothesis of bias does not entail the results of any given number of trials. But it does entail that something such as an improbably extended succession of heads will result over many trials. A hypothesis empirically equivalent to that of bias, were there one, would entail the same thing, and thereby be supported by any evidence thus exemplifying bias. On the other hand, the hypothesis of bias readily admits of evidential support from sources outside its consequence class that would not support purportedly equivalent hypotheses. An example is the information that the coin hypothesized to be biased was poured in a die cut by a chronically inebriated diemaker.

What, then, *is* the connection that we claim our examples to establish between nonconsequential evidence and differential support of empirically equivalent theories? We propose the following exemplar. Theoretical hypotheses H_1 and H_2 are empirically equivalent but conceptually distinct. H_1, but not H_2, is derivable from a more general theory T, which also entails another hypothesis H. An empirical consequence e of H is obtained. e supports H and thereby T. Thus, e provides indirect evidential warrant for H_1, of which it is not a consequence, without affecting the credentials of H_2. Thus, one of two empirically equivalent hypotheses or theories can be evidentially supported to the exclusion of the other by being incorporated into an independently supported, more general theory that does not support the other, although it does predict all the empirical consequences of the other. The view that assimilates evidence to consequences cannot, on pain of incoherence, accept the intuitive, uncontroversial principle that evidential support flows "downward" across the entailment relation.

We stress, however, that this is *only* an exemplar. There are modes of nonconsequential empirical support in science that do not invoke intermediate theories or generalizations. This has already been seen in the case of singular inference, and is present as well in the example from astronomy. Another type of example is the use of analogical reasoning. Analogical reasoning is often motivational, pertaining more to the heuristics of theory development than to confirmation. But sophisticated analogies can be evidentially probative.

Maxwell analogized a closed system of elastic particles to a contained gas, whereby the mathematical theory of collisions together with the observed properties of gases supports a molecular structure for gas. The power of the analogy to yield important, known properties of gases makes it reasonable to infer, Maxwell[14] thinks, that "minute parts" of gases are in rapid motion. Einstein supported his hypothesis of a quantum structure for radiation by analogizing the entropy of monochromatic radiation to the entropy of an ideal gas. He showed that a decrease in entropy corresponding to a reduction in volume for radiation of fixed energy has the same functional form as the decrease in entropy associated with a reduction in volume for an ideal gas. Evidence that warrants treating the gas statistically thereby supports a quantum structure for the gas. Yet such evidence sustains no logical relation to Einstein's hypothesis, nor does it support any theory that entails Einstein's hypothesis.

The many examples we have adduced may seem too obvious and frequent to dwell on. We welcome that reaction, but admonish many influential epistemological positions of the last half century for ignoring them. Karl Popper, for instance, presumes throughout his treatment of scientific methodology that evidence potentially relevant to the assessment of a hypothesis must come from its empirical consequence class. And many epistemologists follow Quine in supposing that the empirical equivalence of rival hypotheses renders inevitable their epistemic parity, reducing a choice among them to purely pragmatic considerations. In defiance of much celebrated epistemology, we claim that our examples establish that theories with exactly the same empirical consequences may admit of differing degrees of evidential support.

B. Empirical consequences that are not evidential. Establishing that evidential results need not be consequences is already enough to block the inference from empirical equivalence to underdetermination. But it is instructive to make the converse point as well. Suppose a televangelist recommends regular reading of scripture to induce puberty in young males. As evidence for his hypothesis (H) that such readings are efficacious, he cites a longitudinal study of 1000 males in Lynchburg, Virginia, who from the age of seven years were forced to read scripture for nine years. Medical examinations after nine years definitively established that all the subjects were pubescent by age sixteen. The putatively evidential statements supplied by the examinations are positive instances of H. But no one other than a resident of Lynchburg, or the like-minded, is likely to grant that the results support H.

---

[14] "Illustrations of the Dynamical Theory of Gases," in *The Scientific Papers of James Clerk Maxwell,* W. D. Niven, ed. (New York: Cambridge, 1890).

This example has a self-serving aspect. That the televangelist has a pro-attitude toward H on grounds independent of the purported evidence he cites is already enough to make one wary; one need not recognize the flaws in the experimental design of the longitudinal study. In a case without this feature, a person hypothesizes that coffee is effective as a remedy for the common cold, having been convinced by finding that colds dissipate after several days of drinking coffee. The point here is that the very idea of experimental controls arises only because we recognize independently that empirical consequences need not be evidential; we recognize independently the need for additional conditions on evidence.

No philosopher of science is willing to grant evidential status to a result *e* with respect to a hypothesis H just because *e* is a consequence of H. That is the point of two centuries of debate over such issues as the independence of *e*, the purpose for which H was introduced, the additional uses to which H may be put, the relation of H to other theories, and so forth.

Results that test a theory and results that are obtainable as empirical consequences of the theory constitute partially nonoverlapping sets. Being an empirical consequence of a theory is neither necessary nor sufficient to qualify a statement as providing evidential support for the theory. Because of this, it is illegitimate to infer from the empirical equivalence of theories that they will fare equally in the face of any possible or conceivable evidence. The thesis of underdetermination, at least in so far as it is founded on presumptions about the possibility of empirical equivalence for theories—or "systems of the world"—stands refuted.

III. FORMAL CONSTRAINTS ON EPISTEMOLOGY

If the identification of empirical consequences with evidential support is so implausible, how has it managed to gain such a foothold? We suggest that a more persuasive, less readily dispelled confusion is ultimately responsible. That confusion, as we have intimated, is to misunderstand the relationship between semantics and epistemology, bringing the largely technical and formal machinery of semantics improperly to bear on epistemic issues.

Specifically, we wish to reveal and challenge the widespread—if usually implicit—conviction that epistemic relations are reducible to semantic relations. It is commonly supposed either that truth and meaning conditions just *are* justification conditions, or, at least, that they can be made to double as justification conditions.[15] Either way, epistemology is made the poor relation of a family of interconnections among semantic, syntactic, and epistemic concepts, and is left to make do with tools handed down from semantics. It seems to us that distinctively epistemic issues are left unresolved by such a presumed reduction, and that epistemic theses depending on it—such as the underdetermination thesis—are wrongheaded. We will first explain and illustrate the confusion we have diagnosed, then trace the mistaken assimilation of support to empirical consequences to it.

---

[15] We do not mean to suggest that all epistemologists are guilty of conflating epistemic and semantic relations. But there are some influential theorists of knowledge, e.g., Quine, who do precisely this.

The problem originates in foundationalist epistemology—especially in Descartes's image of a mathematically rigorous, deductive structure for knowledge—and thus is not confined to empiricism. If the evidential relation is deductive, the evidence on which a knowledge claim is based must bear semantic relations to the claim sufficient to permit the deduction.

Perhaps the best modern illustration is the attempt by the logical empiricists to demarcate science by semantic means. Both "verifiability" and "falsifiability," the prime concepts in terms of which scientific status has been delimited, are tools of semantics. Demarcation criteria proposed by the Vienna Circle or its positivist disciples, and by the Popperians, are alike in depending on semantic analysis and syntactic form of statements. What is required for classification as verifiable or falsifiable is basically that a statement satisfy constraints as to logical form and be couchable in observation language. It was assumed on all sides that such conditions suffice to identify the class of statements that are properly the objects of scientific inquiry.

The problem that demarcation criteria were intended and offered to solve, however, was the epistemic one of judging the reasonableness of belief. What was called for was a distinction, impossible within semantics alone, between statements that are well-founded and those that are not. Whether one's target was the metaphysics, religion, or ethics that exercised the positivist, or the psychoanalysis, Marxism, or astrology that exercised the Popperian, it was the irrationality of credence that the target was to be convicted of, not impropriety of logical form or visual inaccessibility of subject matter.

One might think to defend the adequacy of semantic tools for the intended distinction by arguing that the relevant notion of "science" to be demarcated is not that of what passes muster by scientific standards, but merely that of what is up for grabs in scientific inquiry. After all, statements falsified by scientific inquiry are yet to be classed as scientific. But not only is a distinction between what *qualifies* as scientific and what does not basically epistemic; so too is a distinction between what is *worthy* of investigation or *entertainable* by scientific means and what is not. It is basically what we have already found it reasonable to believe that decides these things.

The demarcation problem of the logical empiricists arose as a variant on the logical positivists' program for distinguishing cognitive significance from emotive uses of language misleadingly given propositional form. Already at this level one may discern the assimilation of evidential to semantic relations. For the evaluative force of, e.g., ethical pronouncements that led positivists to disqualify them as genuine propositions is also present in epistemic pronouncements, and, derivatively, in science. Epistemology, it is now commonly recognized, is value-laden.[16] But science was the logical positivists' paradigm of cognitive significance; its propositional status, the ideal to which ethics, religion, and metaphysics futilely aspired. If epistemology, and science in particular, was to be salvaged, then epistemic evaluation would have to rest on semantic relations as the only factual alternative to value-free empirical relations.

Accounts of scientific explanation and confirmation proposed in the 1940s and 1950s exhibit the same priority of the semantic over the epistemic. Characteristically, these accounts dealt incidently if at all with epistemic and pragmatic dimensions of confirmation and explanation, in favor of their syntax, logical structure, and semantics. Significantly, truth enters as a precondition of explanatory status, rather than as an attribute that it is rational to adduce (partly) in measure that explanations are achieved.

What made such approaches seem plausible, we suggest, was a linguistic view of conceptual analysis. To analyze a concept, one examined its use in language. To understand instances of its use in language, one identified the truth conditions for sentences. In particular, to analyze knowledge one identified the truth conditions for attributions of knowledge; for sentences of the form ". . . knows . . . ." It turns out, however, that among the truth conditions for such a sentence is the truth of another sentence, that disquoted in the second position of the schema. It therefore looks like the semantic concept of truth is logically prior to the epistemic concept of knowledge.

Of course, this does not make the semantic concept sufficient for the epistemic one, but further developments tended to elevate semantics and syntax over the notions of evidential warrant and rationality of belief used in other truth conditions for knowledge attributions. The incompleteness of the list of truth conditions was manifested in a curious asymmetry between the truth and evidence conditions. If the truth condition is not met, no bolstering of the evidence is sufficient for knowing. But inconclusive evidence that leaves open the possibility of error can be sufficient for knowing, if only, as a matter of fact (or happenstance), the world cooperates. In many celebrated paradigms of knowing, the evidence needed does not seem all that strong.[17] Thus, attention focused more on the truth condition than the evidence condition—more on semantic than epistemic issues. Ironically, the recent emergence of reliability theory, which re-emphasizes the justificatory component of knowledge in the tradition of Gettier's challenge, underscores the paucity and defeasibility of the evidence on which ordinary knowledge relies. Add to this asymmetry the success of Tarski's theory of truth in contrast to the sorry state of theories of evidential warrant, and one has the makings of a semantic and syntactic orientation for epistemology.

---

[16] See Lycan, "Epistemic Value," *Synthese*, LXIV (1985): 137–64.

Given this orientation, it was natural to approach the problem of warranting a hypothesis—the problem of testing—by attending to statements that bear syntactic and semantic relations to the hypothesis—to its instantiations. At least this was natural for empirical generalizations, whose instantiations are empirical statements. This approach then created so many internal problems and tasks—Hempel's paradoxes of confirmation across logical relations; Goodman's problem of projectability—that the possibility of warrant provided by statements syntactically and semantically independent of the hypothesis was lost sight of. Instantiations of theoretical hypotheses are not empirical, but an assimilation of support to consequences was somehow extrapolated for them, by supposing them in principle recastable in observational terms or, perhaps, by supposing their testability reducible to the testability of empirical generalizations. Such was the hold of the resulting picture, that the assimilation of support to consequences exceeded the confines of logical empiricism to capture the format of textbook characterizations of scientific method itself. Although written by a philosopher, Hempel,[18] the following passage will strike every reader as stereotypical of standard accounts of empirical inquiry:

> First, from the hypothesis under test, suitable other statements are inferred which describe certain directly observable phenomena that should be found to occur under specifiable circumstances if the hy-

---

[17] This is a striking feature of Gilbert Harmon's examples in *Thought* (Princeton: University Press, 1973).

[18] *Aspects of Scientific Explanation* (New York: Free Press, 1965).

> pothesis is true; then those inferred statements are tested directly, i.e., by checking whether the specified phenomena do in fact occur; finally, the proposed hypothesis is accepted or rejected in the light of the outcomes of those tests (*ibid.*, p. 83).

This is a pure and simple statement of a view that deserves to be called "consequentialism," viz., the thesis that hypotheses are to be tested exclusively by an exploration of the truth status of those empirically decidable statements which they entail. Consequentialism is closely connected historically with the ancient idea of "saving the phenomena," that is, of reconciling theory with aberrant phenomena by introducing auxiliaries permitting those phenomena to be derived. Although not the invention of empiricists, it held a special appeal for them in appearing to solve problems of meaning, truth, and justification in one go. Meaning conditions, truth conditions, and justification conditions became substantially the same thing. The pre-Tarskian slogan, "To know what *x* means is to know how to test *x*," became, after Tarski, "The truth conditions for *x* specify simultaneously the meaning of *x* and the test conditions for *x*." [19] And this cast of mind, we suggest, is behind the supposition that statements with the same empirical consequences must be on an equivalent epistemological footing.

It is remarkable, upon reflection, how much freight the entailment relation has been made to carry in recent philosophy. The early logical positivists identified the meaning of a statement with what that statement entailed in the observation language. The logical empiricists and Popper sought to demarcate scientific from nonscientific statements solely by whether or not they entailed observational statements (Rudolf Carnap) or negations of observation statements (Popper). The Tarskian account of truth held that a theory was true just in case every statement constituting its logical content was true, and in practice a statement's logical content was to be fixed by its entailments.

In twentieth-century epistemology, we find the notorious Nicod criterion, discussed at length by Hempel and other confirmation theorists of the 1950s and 1960s, presuming to define the evidential relation exclusively in terms of positive and negative instances of the hypothesis, where positive instances instantiate the hypothesis and negative instances instantiate its negation. Popper's influential account of theory testing held that genuine tests of a theory had to be drawn from statements in its empirical content class, a class defined as the set of empirical statements whose negations were entailed by the theory. Currently, van Fraassen's constructive empiricism defines its key evaluative concept—empirical adequacy—in terms of the observational structures that model a theory:

---

[19] Compare Quine: ". . . epistemology now becomes semantics. For epistemology remains centered as always on evidence; and meaning remains centered as always on verification; and evidence is verification"—"Epistemology Naturalized," p. 89.

> . . . a theory is empirically adequate exactly if what it says about the observable things and events in this world, is true—exactly, if it 'saves the phenomena' (*op. cit.*, p. 12).

And that a structure models a theory is determined not, as we have remarked, definitionally, but by reference to the theory's entailments. Clark Glymour, an avowed scientific realist, develops a "boot-strapping" story about theory testing, which insists that the statements capable of testing a hypothesis imply its empirical consequences. And Quine repeatedly avers that the only central rule of scientific method is hypothetico-deduction.

This ubiquitous assimilation of a theory's test cases to its logical consequences in an observation language, as we have argued above, wrongly ignores some of the more salient ways of testing theories. Worse, it generously greases the slide from empirical equivalence to underdetermination and epistemic parity. Ironically, the limitation of a statement's justification conditions to its truth conditions represents a striking break with the traditional empiricist project. Prior to the emergence of neopositivism in the 1920s, the general idea about theory testing and evaluation was that there was a range of "phenomena" for which any theory in a particular field was epistemically accountable. (In planetary astronomy, for example, these phenomena would be observations of positions of the planets, sun, and moon.) A theory's success or failure was measured against these phenomena, and decided by the theory's ability to give an account of them. A theory was, of course, responsible for its entailments, but it was held equally accountable for all the relevant, established phenomena, and could not evade this responsibility by failing to address them. For a Newton, a Ptolemy, or a Mach, "saving the phenomena" meant being able to explain all the salient facts in the relevant domain.[20]

With the rise of neopositivism, the epistemic responsibilities of theories were radically reinterpreted. Theories became liable only for what they entailed. Failure to address relevant phenomena, or at least to be indirectly applicable to them, now emerges as a cheap way of protecting such success as a theory does achieve, rather than as a liability. Where empirical adequacy formerly meant the ability to explain and predict all the salient phenomena, it now requires only possession of none but true empirical consequences. Recall the passage lately quoted from van Fraassen. The radical character of the shift we are describing becomes immediately clear there when one notes his identification of "empirical adequacy," saying only true things about observable features of the world, and "saving the phenomena." Prior to our time, no one would have supposed, as does van Fraassen, that saving the phenomena amounts only to possessing an observable model. No one would have supposed, as does van Fraassen, that a theory is to be judged only against the correctness of its own observational commitments (be those commitments expressed in model-theoretic or propositional form), irrespective of the comprehensiveness of the class of such commitments, irrespective of the theory's applicability to problems independently raised. It is testimony to the pervasiveness of the thesis that epistemic assessment is reducible to semantics that van Fraassen's conflation of the hitherto quite disparate notions of empirical adequacy and saving the phenomena has gone unnoted.

---

[20] As Duhem, summarizing the state of instrumentalism in 1908, put it: "we now require that . . . [scientific theories] save all the phenomena of the inanimate universe together"—*To Save the Phenomena* (Chicago: University Press, 1969), p. 117.

Much epistemology in our day is arbitrarily and unreasonably constrained by these developments. Our concluding, positive moral is that epistemic warrant unfettered by semantics has rich and varied sources yet to be exploited.

LARRY LAUDAN
University of Hawaii/Manoa

JARRETT LEPLIN
University of North Carolina/Greensboro

PENNY PINCHING AND BACKWARD INDUCTION*

"REASON is the pace, increase of science the way and the benefit of mankind the end." Hobbes makes this remark early in *Leviathan*, meaning it to apply both to the study of human behavior and to human behavior itself. Reason bids each of us do what benefits himself; science helps each to identify the rational choices; the results benefit all. The modern science most deeply committed to this hope is neo-Classical economics, and its most enlightening analysis of reason is supplied by game theory. The general idea is to study economic behavior as choices made by rational agents in allocating scarce resources. In so far as real-world choices are indeed rational, the science is descriptive. In so far as they are not, it is a normative science advising us how to fare better. In either case, the yardsticks are established by exploring an ideal-type world, where all choices are rational choices, in a sense defined by the theory of rational choice and put to work by game theory.

Notoriously, it soon emerges that individually rational choices need not sum to collectively rational results. Here is an example. *A* and *B* play a game where ten pennies (kindly donated by a third party) are put on the table and each in turn takes one penny or two. If one is taken, the turn passes but as soon as two are taken the game stops and any remaining pennies vanish. What will happen, if both players are rational? Offhand one might suppose that they emerge with five pennies each, or with a six-four split because the player with the odd-numbered turns takes two at the end. But game theory seems to say not. Its apparent answer is that the opening player will take two pennies, thus killing the golden goose at the start and leaving both worse off.

The immediate trouble is caused by what has become known as backward induction. Figure 1 sets the game out as a series of choices between moving "across" by taking one penny and moving "down" by taking two (and thus ending the game). The resulting pennies gained by each player are given by the bracketed numbers, with *A*'s put first in each case. Looking ahead, *B* realizes that they will not reach (5,5), because *A* would settle for (6,4). *A* realizes that *B* would therefore settle for (4,5), which makes it rational for *A* to stop at

---

* This paper is part of the Foundations of Rational Choice Theory project in progress at the University of East Anglia with the support of the Economic and Social Research Council (award R 000 23 2269). I would like to thank Robin Cubitt, Shaun Hargreaves Heap, Adam Morton, Judith Mehta, Chris Starmer, and Robert Sugden for their helpful comments on a previous draft.

0022-362X/91/8809/473–88

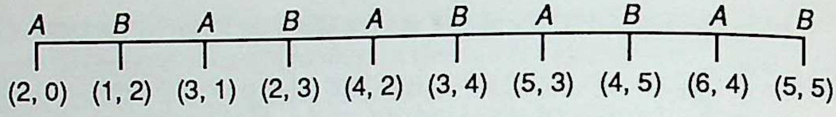

Figure 1

(5,3). In that case, *B* would settle for (3,4); so *A* would therefore settle for (4,2), leading *B* to prefer (2,3); and so on. *A* thus opens by playing down and reason has obstructed the benefit of mankind.

One could try shrugging off this disconcerting result on the grounds that there is nothing odd about it, since, as with the one-shot or finite prisoner's dilemma, rational choices need not be Pareto-efficient in sum. But it is certainly disconcerting. With, say, a hundred pennies on the table, one would swear that rational players would manage to take at least, say, forty-five each before being unnerved by cat-and-mouse considerations, as indeed ordinary people commonly do in experiments with such games. Yet the theoretical proof that (2,0) is the outcome of a game of any finite known length between rational agents is exact and robust, as I shall try to show. At the same time, it threatens catastrophe for both the descriptive and the normative claims of neo-Classical economics, as I shall also seek to show. Yet all is not lost, since there are ways of avoiding the elenchus, if the standard assumptions of game theory are challenged at a finely calculated depth.

I. STANDARD ASSUMPTIONS

Consider a simpler series with only five pennies on the table. The stepping stones of the backward induction are then these:

(1) *B* would take two pennies at his second turn.
(2) *A* would take two pennies at his second turn.
(3) *B* would take two pennies at his first turn.
(4) *A* will take two pennies at his first turn.

To justify and connect them, the assumptions required for a game between ideally rational agents can be conveniently specified thus:

(M) *Motivation*. Each player has complete and consistent preferences for the possible outcomes, ordered on the principle that more for himself is better than less.

(I) *Information*. Each knows the rules and conditions, and can draw all warranted inferences. This includes knowing that both players are motivated as above and are rational in a sense given by the three conditions together.

(K) *Common Knowledge*. Each knows whatever the other knows. (Thus, if *A* knows that *p*, then *A* knows that *B* knows that *A* knows . . . etc. . . . that *p*.)

Given these three assumptions, each stepping stone seems secure and the backward induction is therefore at least tempting.

A preliminary query concerns (M): if the assumptions are meant to specify a notion of rationality, it is not plain that they should make rational agents self-interested in the sense given by (M). For the moment, however, I postpone this topic with the preliminary remark that the players cannot make rational choices without some mutual knowledge of what is motivating each of them. (M) is convenient both because it is a standard assumption and because it helps to spark the backward induction neatly. It does the symptomatic job of giving each player an aim that seems reasonable in itself but oddly self-defeating when the other player has a matching aim.

It may also be asked why I have separated (I), mutual information about the rules of the game, the pay-offs, and the players, from (K), common knowledge. (I) is meant to make the players translucent, so to speak, in ways that are uncontentious, while also raising a question about how far translucency can extend. (K) offers the answer that ideally rational agents are fully transparent to one another, thus clearing the way for the suggestion that the source of the trouble with backward induction lies in the peculiarities of this limiting case.

II. COMMON KNOWLEDGE

Common knowledge, sweepingly defined so that "each knows whatever the other knows," is a condition so strong that it may make it *impossible* for the opening player to take only one penny and hence *impossible* that any later node in figure 1 will be reached. Hence one might protest that it debars *A* from reasoning about what would happen were the game to proceed, since the reasoning would rely on premises that he knows to be false. In that case, the assumptions that generate the paradox would need restating and the question would become whether, as Jon Elster[1] puts it, "there exists a set of assumptions about the players which are strong enough to allow us to infer what the players would do at various nodes of the game while also weak enough to be consistent with those nodes being reached" (*ibid.*, p. 5).[2]

Philip Pettit and Robert Sugden[3] propose a restatement of the common knowledge assumption that disposes of the backward induction paradox. They start by arguing that "under the assumption of common knowledge neither player is allowed to think strategically" and hence that the paradox cannot get started. Translated for the penny game, the nub of their argument is this. Given (K), *A* must open by taking two pennies. For were he to take only one, *B* contemplating this "pass" and applying backward induction, would conclude that *A* had acted irrationally. But *B knows* that *A* is rational. Hence there is a coherent tale to tell only if *B* starts merely with a *belief* that *A* is rational and can be led by an opening pass to revise this belief. Equally, *A* can coherently contemplate later nodes being reached only if he at first *believes*, but does not *know*, that *B* is rational, and stands ready to revise that belief, if *B* were to pass. But, when common knowledge is replaced with revisable common belief, so that backward induction has possible outcomes to work on, it ceases to be uniquely rational to open by taking two pennies and the backward induction is thwarted. Therefore, the paradox either does not arise (common knowledge) or is resolved (common belief).

---

[1] *The Cement of Society* (New York: Cambridge, 1989).

[2] Elster sets the problem out crisply and cites an unpublished paper by Christina Biccieri. See also K. Binmore, "Modelling Rational Players," *Economics and Philosophy*, III, 4 (1987): 179–214, for part I; IV, 1 (1988): 9–55, for part II.

[3] "The Backward Induction Paradox," this JOURNAL, LXXXVI, 4 (1989): 169–82. In what follows, I shall take my cue chiefly from their fine paper, which I recommend warmly even while disputing its thesis.

I agree that complete mutual transparency is a highly peculiar condition, even for an ideal-type, limiting case. But the oddity does not lie in the logic of the concept of knowledge. Pettit and Sugden claim that, under the assumption of common knowledge, both players in a one-shot prisoner's dilemma "*must* defect" (*ibid.*, p. 181). 'Must' is too strong. Even if they *will* defect, given common knowledge, it does not follow that they *must*. "Necessarily, if *S* knows that *p*, then *p*" does not imply "If *S* knows that *p*, then necessarily *p*." When gods play chess, each knows that the other *will* make no mistakes and, in this knowledge, identifies inferior moves by considering what would certainly happen if they were made. This *reductio ad absurdum* reasoning proceeds by contemplating positions that will not be reached. It could not proceed if such positions were impossible. But they do not become impossible as a consequence of the players' being omniscient.

The peculiarity of common knowledge seems to me to lie not in its logic but in the sheer transparency that it confers on the players. Any reason that might move *A* to pass would be known to *B* and vice versa. Any traps and strategems would be instantly detected. This would certainly take the steam out of zero-sum games like chess. But even common belief does the same for simpler ones like tic-tac-toe, where humanly rational players always draw. The penny game, however, like the prisoner's dilemma, is not zero-sum. Both players can do better than play safe if and only if *both* players will rationally risk it. Pettit and Sugden are, I think, right to say that it can be rational to contribute to the better outcome only if the players are somewhat opaque to one another. But it is not plain that the doubt thus licensed operates by shaking an initial belief that the other player is rational.

The general idea is that an opening pass by *A* would inject an indeterminacy whose upshot is that *A* no longer has a uniquely rational opening move. I can see three ways to make the case but find none of them convincing. Firstly, an opening pass might lead *B* to judge *A* utterly irrational, thus withdrawing a premise needed to run the backward induction and so letting *B* pass in the hope of a further pass by *A*. Taken in isolation, however, this dissolves all scope for reasoning ahead, since *B* has to judge *A* to be extremely stupid. Otherwise, since the players are still motivated (M) to prefer more to less at the end, *A* would still opt for (6-4) rather than (5-5) and *B* would still pre-empt this. Unless *A* is demented, this is enough to activate the backward induction.

Secondly, then, an opening pass might lead *B* to credit *A* with rationally permissible beliefs that would warrant a pass. These beliefs must include a belief that *B* will pass in response but can vary in the reasons that *A* attributes to *B*. Thus, *A* might expect *B* to judge *A* demented, as above, or might expect *B* to conjecture that *A* is following one of various rules, each of which involves at least one further pass by *A* (although differing in how many). Provided that *B* has several possible conclusions and is unsure whether or how completely *A* is rational, then this prospective uncertainty can enter *A*'s original calculation. Unfortunately, however, the previous snag stands. If *A*'s plan is well laid, *B* puts a stop to it by killing the game one move before *A* intends, with backward induction ensuring that this means at once. *B* does the same, if *A* is irrational, as above. Hence the second option leads nowhere.

Thirdly and most plausibly, *A*'s opening pass might show him to hold rationally permissible beliefs that neither forbid a pass nor require one. They must include a belief that *B* may (not will) pass in response. This belief by *A* is warranted because indeterminacy breeds indeterminacy not only in *B*'s mind but also, through feedback, in the mind of *A* doing the original calculation. This amplifies the curious element of self-reference in the expectations of each player to the point where the backward induction loses its grip and each is warranted in whatever he does. Paradox is met with paradox, thus inviting analogies with, for instance, the surprise examination, where the pupils' *reductio* of the teacher's promise seems to reinstate the promise. That sounds like the right sort of place to locate the trouble.

But the cure seems to me too drastic, if the price of stopping the backward induction is to leave rational agents without a rational course of action. Game theory might do better to hang on to the disease! Consider the closing stages of the penny game, if indeterminacy is injected into the opening stages. Would *A* still take two pennies, when the only remaining outcomes are (6, 4) and (5, 5)? Would *B* still pre-empt him at the turn before, so that *A* would still cash in at the turn before that? Pettit and Sugden presumably mean indeterminacy to reign here, too. In that case, consider the very simple game (of the penny sort but with slightly different rules and pay-offs) in figure 2. Could *A* pass? The standard answer is that he could but, being rational, will not. If *A* were to pass, would *B* have a determinate response? The standard answer is that *B* would play down, because he prefers (0,3) to (2,2). To reject these answers is to destroy the roots of game theory.

Meanwhile, there is a less destructive line to pursue, and I adjourn further discussion of reason until more has been said about (M). That more pennies are better than less has, so far, prompted *A* in figure 1 to settle for (2,0), because he prefers (6,4) to (5,5), thus somehow losing sight of the fact that (4,5) would have been better for him than (2,0). Can we restore his common sense by a challenge to the standard game-theory idea of what motivates rational players?

III. MOTIVATION

Game theory is a formal system in which a "game" is a set of choices by instrumentally rational agents, where the pay-off to each depends on the choice made by others. This system seems to have quirks, as when rational choices sum to outcomes inferior for all. The quirks are teasing because such "games" invite thought about the games of flesh-and-blood social life. Thus, the penny game between *A* and *B* illuminates the self-defeating character of some everyday encounters between, one might say, Jack and Jill. At the same time, however, one notices that Jack and Jill do not always land themselves with the disappointments of *A* and *B*. Is the difference simply that *A* and *B* are more rational than Jack and Jill, or has the style of abstraction filtered out something crucial to social interaction? With this question in mind, I turn to motivation.

In the analysis of the penny game, motivation is conceived in Humean terms. Reason alone cannot be a motive to any action of the will; it is the slave of the passions and its only office is to serve and obey them. Whether *A* and *B* choose rationally depends not on

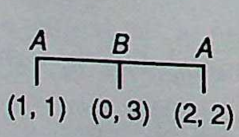

Figure 2

what their passions are (provided that they are consistent) but on how well their choices satisfy their passions. To perform the calculations needed, however, there has to be a unifying measure of satisfaction, standardly called "utility." Although conceived originally as a psychic flow or inner glow, utility now merely marks the presence of a preference; but the claim to unity is preserved by specifying complete and consistent preferences in a rational agent. By keeping silent about sources of utility, moreover, a universal analysis of motivation can be offered, applying equally to egoists and altruists, patriots and cowards, Christians and atheists. Every rational agent is then wholly self-interested, in the very precise sense that he endeavors to maximize his utility.

This means that *A* and *B* play the penny game not for pennies but for what we can call, with seemingly harmless poetic license, utils. Jack and Jill, eyeing figure 1 and thinking in pennies, might not care about one less or more or might actively prefer a fair outcome of five each. *A* and *B* have no such room for maneuver. For them figure 1 gives pay-offs in utils, which are determined in advance of play by taking any and every source of satisfaction into account. The numbers shown represent the given preferences of *A* and *B*, scored in utils, over the outcomes in pennies. In the present case, the utils and pennies happen to match. Were it not so, then the utility numbers would be the ones that mattered.

Utility numbers thus have a dual function. They serve as information available to both players and they also motivate rational choices. The motivating function is, so far at least, wholly automatic. If the game ever reached the penultimate node, *A* would automatically take two pennies, thus securing himself a (6,4) outcome. This, I submit, is all it takes to animate the backward induction. One might have hoped that reflection could rescue them by pointing out the mutual benefits of an indirect strategy, whereby each distances himself from the utility numbers and acts as if upon a different set of figures. But there is simply no scope, while the numbers have their dual function.

That may seem too quick. It is certainly too quick for the godfather of this analysis of motivation. Hume remarks in his essay, "Of the Original Contract":[4]

> For, as it is evident that every man loves himself better than every other person, he is naturally impelled to extend his acquisitions as much as possible; and nothing can restrain him in this propensity but reflection

---

[4] In E. Barker, ed., *Social Contract: Essays by Locke, Hume, and Rousseau* (New York: Oxford, 1956).

> and experience, by which he learns the pernicious effects of that licence, and the total dissolution of society which must ensue from it (*ibid.*, p. 277).

He credits "reflection and experience" with the power to prevent otherwise pernicious effects, helped along by the fact that self-love is not our sole motivating sentiment. Adam Smith works the same street in *The Theory of Moral Sentiments,*[5] suggesting at one point that we have something of an impartial spectator in us—enough to secure "the harmony of society":

> However selfish mankind may be supposed, there are evidently some principles in his nature which interest him in the fortunes of others . . . . By imagination we place ourselves in his situation (*ibid.*, p. 9).

But, once utility numbers have been assigned, the sentiments that occasioned them have no further part to play. If reason's only task is to point the agent toward the larger number, then reflection and experience are powerless and the impartial spectator can only weep.

This may still seem too quick. Even if some sentiments, like the natural sympathy that Hume thought a crucial element in human nature, affect only the assignment of utilities to outcomes, others, like a willingness to view ourselves impartially, might intervene between utilities and actions. After all, the utilities in the penny game are not *expected* utilities, as is all too plain if the expectation of reaching the later nodes is nil! So can "reflection and experience" not stop a rational agent applying the principle that more for himself is better than less so as to emerge with less? Reflection notices that it would be rational for *A* to start by passing if and only if he expected *B*, moved by parallel reflection, to pass. In many similar games, there are several sets of reciprocal expectations that offer the players a grip on each other's bootstraps. In the penny game, is there not enough of an interlock of expectations to let reflection coax *A* and *B* past the first obstacle?

I fear not. So far, *A* is so motivated that he can be relied on to take two pennies if only two remain, and *B* is so motivated as to deny him the chance. An initial pass by *A* could affect these reliable expectations only if it could count as a credible commitment or threat. But, since there is no scope for reprisals and no way for a player to burn his boats and shut off the troublesome endplay, it remains foolish to cooperate. Credibility implies that both players can rationally rely on the game ending with (5,5) as a result of rational choices; and neither *A* nor, therefore, *B* has any reason to expect this amiable outcome.

---

[5] New York: Oxford, 1976.

That narrow-minded retort, however, depends on a further feature of the motivational assumption. Implicitly the scene has been so set that only forward-looking reasons can motivate. Commitments, and indeed threats, involve backward-looking reasons. If Jack makes Jill a promise, he undertakes to let the fact of having promised move him at a later date, even though a forward calculation at that point might show it more profitable for him not to. To threaten Jill credibly, he must convince her that her having done what he had warned against doing will prompt him to retaliate even at a loss. None of this apparatus is so far available to *A* and *B*. What *A* can expect *B* to do, and vice versa, depends solely on what will lie ahead at the point of decision, as if each turn were the start of a new game. Even if *A* is allowed to utter the words of a promise or threat, the words change nothing, because *A* will do what he would have done without them, namely, make the most rational move and let bygones be bygones.

The forward-looking character of reasons in game theory is less obtrusive for repeated games. Where there is more of a future to play for, it may make sense to invest in a reputation for constancy even at some immediate cost. But the reasons for doing so are still forward-looking and they carry conviction only by reference to improved prospects. Even in supergames no one does what he has promised solely for the reason that he has promised, and, in so far as the concept of promising requires that having promised can be a sole reason, supergame practices are borrowing from a world beyond game theory. Similarly, the "ethical" preferences sometimes invoked to account for otherwise irrational choices can work only by affecting the utility numbers before the game starts. In the analysis provided so far, they are simply an element in a preference set that is complete, consistent, and a matter of public information. There is no scope for giving them special influence during play, once the utility of outcomes has been established in advance.

A Rawlsian veil of ignorance sounds promising here. If *A* and *B* know everything about the game, its pay-off, and the forward-looking reasons that they will activate, but do not know who will have first move, then they might see things through the eyes of an ideal spectator. It is as if they were to play the game with *A* starting but then to allot the pay-offs afterward by tossing for who got which. Would they not rationally agree that cooperation leading to (5,5) is the best insurance policy and hence that whoever turns out to be "first player" will pass? Yes, they have forward-looking reasons to assent. But ignorance has not turned egoists into Kantians or even utilitarians, and as soon as the veil is lifted there will be forward-looking reasons not to comply.

The general snag is that it simply does not matter how exactly the moral psychology of the agents works, while utility numbers are given in advance, preferences are complete and consistent, and these two sides of a single coin carry the players irresistably forward. While the utilities have their dual function, otherwise promising thoughts about other sorts of reasons and two-tier moral psychologies where reflection can make a difference are frustrated. Yet, to break their grip, we must stand further back.

IV. REASONS AND MOTIVES

I conclude so far that *A* and *B* are stuck with the backward induction paradox, which needs only fixed pay-offs and an assumption that motivation is supplied by forward-looking reasons addressed to the pay-offs. Yet I do not draw this conclusion for Jack and Jill and the human encounters that game theory idealizes. Try to think of a game of social life that is played by people who know as much and as little about one another as *A* and *B* do. The international stock exchange? It could not function without trust and, hence, even if one is duly cynical, without screening devices to ensure participants who take normative expectations seriously when deciding what to do next. Freak games between complete strangers, as when a lifeboat after a shipwreck has only a tiny stock of water? Disasters do not create a social vacuum large enough to destroy all normative expectations. How about games across the garden fence, where neighborly bonds are only skin deep, as when you and I jointly cultivate a superb melon in the uneasy knowledge that, if it disappears one dark night, the other can get away with blaming vandals? We would never have started, had either suspected the other of being a game theorist!

The basis of economics is often said to be exchange—an interesting idea, if one warms to George Simmel's dictum that "all contacts among men rest on the schema of giving and returning the equivalence." In an exchange, according to game theory, each party performs an action that is rational and the exchange is the sum of these separate choices. The penny game looks at first like an exchange made in a series of turns, each of which is a rational choice. The backward induction puts pressure on what is supposed to connect the series. When Jack and Jill play the penny game, Jack opens by passing, thus making Jill an offer, which Jill accepts by passing, too, and Jack honors his obligation by passing again. How foolish of them, if reasons operate only forward! Yet, in that case, they do not operate at all and Jack becomes *A*, who is led by backward induction to turn down the chance of more.

The contrast between kinds of reason may be clearer with a change of example. Suppose that Aunt Maud passes a hat around to her nephews Tom, Dick, and Harry, promising to double whatever money they put in and then to share the proceeds equally among them. If Tom is a well-tutored game theorist, he analyzes the conditions as a game between *A*, *B*, and *C*, where each invests x pennies and gets back two-thirds of his investment plus two-thirds of everyone else's. Thinking as *A*, he notes that he gets two-thirds of the rest whether he invests anything or not, and makes a loss on anything invested by himself; so he puts nothing into the hat. Perhaps he stops to ask whether a contribution might stimulate one by *B*. But he then reflects that *B* will reason as he has just done, regardless of what is already in the hat, and so will *C*. Aunt Maud can be confident that her amiable gesture will cost her nothing.

If Tom is not so well-tutored, however, fraternal relations should produce a different outcome. Admittedly, it is not plain how much he is to contribute. What if he has 50 pence in his piggy bank, Dick £1 and Harry 10 pence? What if he is the oldest? But, even if the notion of fair play among brothers is ambiguous, contributions of, say, $50 + 75 + 10$ will produce 90 pence each and so a profit for all. More simply, if Aunt Maud puts £3 in the hat before passing it round, saying that each may take £1, if he wishes, and that she will double whatever is not taken and share it out equally, then, whereas *A*, *B*, and *C* would each make £1, Tom, Dick, and Harry should each make £2.

'Should' occurs twice in the last paragraph and is tricky. One might construe "fraternal relations" as meaning that Tom, Dick, and Harry are morally or socially bound to help produce the largest collective benefit. But that would take us too far afield. I want to propose the weakest reading of the scene which includes kinds of reasons missing from the atomistic game-theory reading. The crux lies in the difference between normative and rational expectations. So far we have been dealing with the latter. When the players are *A*, *B*, and *C*, what each expects is what each predicts. Even if *A* can somehow borrow from the stock of social practices to signal for an exchange, he can only predict that *B* and *C* will find no forward-looking reason to respond. Bygones are always bygones. Tom, Dick, and Harry, by contrast, are related by social ties governing what they can expect *of* one another. These expectations are normative in the sense that failure to comply attracts moral criticism, especially from those entitled to expect compliance. A moral vocabulary comes into play, even if one relative to a particular social and cultural setting rather than grounded in pure ethics.

In this vocabulary, exchange involves reciprocal obligation, set within an existing scheme of normative relations. By accepting what is offered, the second party acquires a reason to do his bit which it will take a special justification to ignore. Wittgensteinian thoughts arise about the character of action in relation to rules and suggest a fresh sense in which social interaction consists of "games." Although I cannot pursue them here, I mention Wittgenstein as one way of showing how past actions can give rise to backward-looking reasons. If Dick *owes* Tom £5, then this is a reason for him to give Tom £5 which is distinct from any reason having to do with taking pity on Tom, cheering him up, or making a shrewd investment. Tom expects to receive £5 in two different senses of 'expect'. He expects it of Dick as a matter of entitlement and he predicts that it will be forthcoming. That these are different senses is clear, if one notes that Tom can be owed £5 but predict sadly that Dick is not going to pay up. I mention Wittgenstein to indicate that it takes a profound analysis to spell out these homely points.

We should not regard social actors as clockwork toys in the domain of normative relations. There is no need to construe *homo sociologicus* as a throughput between social structure and social action. The playing of social games takes judgment even when one seeks only to play the game aright. We can credit social actors with the power to step back reflectively within the game of the moment and also to step out of it. Tom, Dick, and Harry illustrate both possibilities. The fact that they are brothers signifies only indefinite fraternal expectations, and the commercial character of Aunt Maud's offer makes them wonder how far brotherhood is relevant. Such ambiguities leave plenty of room for reflection and experience on the part of what are nonetheless social actors. The force of calling them "social," however, remains their openness to backward-looking reasons for action. They have the conceptual makings of an exchange for mutual benefit, even though that does not guarantee success in staging one.

Exchange, like many social or moral transactions, requires trust—a reliable normative expectation that today's forward-looking reasons will become tomorrow's backward-looking reasons. Game theorists and contractarians see this as clearly as anyone. So they are tempted to try building on the thought that there are forward-looking reasons to introduce a scheme of backward-looking reasons, rather as rule-utilitarians try stopping the rot that act-utilitarianism spreads in institutions, by the two-tier device of entrenching the most useful rules. It is as if reflection could lead *homo economicus* to beget *homo sociologicus,* who, having internalized the normative expectations surrounding a transaction, reworks the utility numbers before handing back to *homo economicus* to make the choices.

If indirect solutions of this kind worked, they would have the interesting result of shifting the basis of economics from decision theory to the theory of organizations and institutions. Economic agents would enter markets essentially as members of firms, banks, government agencies, unions, churches, and households, since the serious work would be done in the assigning of utilities. But, first, there is still a problem in seeing how an indirect solution could work.

If one asks exactly why backward induction holds in the penny game, the immediate answer is that *A* and *B* are too cramped by the setting to be able to strike a bargain. Even if *A* signals for one by passing, *B* will ensure against the bargain's not being kept or, more simply, will stand ready to exploit it. This suggests that they might get on better if they were less cramped, for instance, if they could talk the game over or make a bargain in front of witnesses. But such human possibilities require a fresh moral psychology, if the problem of compliance (of giving *A* sufficient reason not to stop the game at (6,4)) is not to drive out trust. It has emerged, moreover, that nothing can be done in this direction unless a rational agent can come to override the apparent *diktat* of the utility numbers. In the absence of a moral psychology that allows all this without merely destroying the basis of game theory, the cramped setting of the penny game is fair comment.

So let me now try to answer Elster's question whether "there exists a set of assumptions about the players which are strong enough to allow us to infer what the players would do at various nodes of the game while also weak enough to be consistent with those nodes being reached." I suggest that we start by stating the pay-offs in pennies after all, so as to be able to break the stranglehold of the utility numbers. Then let us make the ranking of an outcome in a player's order of preference depend not only on how many pennies it brings but also on how it is reached. Thus, six pennies gained by breach of trust need not outrank five gained honorably, or indeed the four that *A* gets by acting decently despite *B*'s intended failure to act likewise. This is no simple amendment, however, since it involves importing a normative vocabulary into the description of the outcomes and also refashioning *A* and *B* so that normative descriptions can make a difference.

Condition (M) therefore needs restating carefully. A first thought is that the players can still have complete and consistent preferences over outcomes in advance of play, if the preferences range over routes as well as destinations. But, if we construe this by analogy with a journey where it is worth something to travel in comfort through pleasant scenery, we shall still be dealing in utilities. That, admittedly, keeps the moment of decision clear. In the penny game, for instance, the more profitable outcomes can only be reached by routes where the players make and keep what are in effect bargains. A shorter version played with a total of six pennies and two players who prefer to play fair might look as in figure 3:

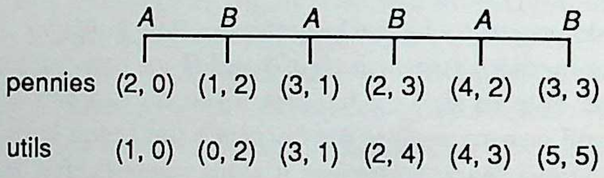
A B A B A B
pennies (2, 0) (1, 2) (3, 1) (2, 3) (4, 2) (3, 3)
utils (1, 0) (0, 2) (3, 1) (2, 4) (4, 3) (5, 5)

Figure 3

Here a forward induction will keep *A* and *B* going to the end. *A* opens with a pass because he knows that *B* will pass because *A* will pass again because *B* will pass again because *A* prefers (5,5) to (4,3), when the pay-offs are stated in utils.

But this is no solution at all. It merely rewrites the penny game arbitrarily so that it no longer illustrates the problem. That the translation into utils is arbitrary is plain, if one considers the intermediate outcomes, where more pennies dishonorably gained have been ranked in comparison with less pennies honorably gained. Meanwhile, nothing has been done to introduce the backward-looking reasons needed for honorable conduct. We need a fresh version of (M) that leaves the players a possible escape when faced with a threat of backward induction.

A very different version might make rational players Kantians, who choose whatever Adam Smith's impartial spectator would recommend, regardless of their individual profit. Thus, to revert to the game in figure 2, *A* plays across and *B*, being a Kantian, too, also plays across, and they emerge both rich and virtuous. But this seems altogether too easy, not least because the backward induction problem has again simply vanished. Kantians always have a dominant strategy of heading for the right outcome, regardless of what other people will do. Two Kantians will therefore achieve it, but there is no warrant for saddling game theory with a motivational assumption that ensures that only Kantians can play with success.

I introduced figure 2 earlier to make the point that radical indeterminacy would destroy the basis of game theory. If *A* could leave *B* without a rational course of action by passing, then the theory of rational choice collapses. But at that stage motivating reasons were only forward-looking and it made no difference how the players came to arrive at figure 2. Similarly, if *A* were to play across, that would create a new game consisting of figure 2 with the left-hand node deleted. What we now need, I suggest, is a version of (M) where it can matter what happened earlier in a game, because the players can have reason to care, but where compliance with bargains is not guaranteed. This suggests rewriting (M) thus:

> (M*) *Motivation:* each player ranks the possible outcomes by the measure of what best serves the interests of both players, provided that he expects the other player to rank them in the same way.

This version lets *A* open the penny game with a pass, because that will encourage *B* to expect a further pass. But the proviso in it is meant to make the players translucent rather than transparent. As with Pettit and Sugden, therefore, there is an unsettling element of mutual self-reference. But they located it in the initial common belief that the other player is rational, where it spreads like dry rot. By locating it in an initial common belief about motivation, I hope to keep it within bounds. For, although the players again have no certain course of action, they are simply short of information which, if they had, would not introduce a further paradox.

This translucency is due in part to the game-theory abstraction that turns Jack and Jill into *A* and *B*, thus shelling them out of their normal and normative social relations. Since *A* and *B* are strangers, *A* has excellent reason to settle for (1,1) in figure 2 and some reason not to risk even one pass in the original penny game. But Jack and Jill are also translucent, even if they know more about each other. They are *persons* and thus inherently better placed to act in the light of "reflection and experience," as Hume recommends, but still separate persons, whose honorable character is not guaranteed. That leaves much more to be said about the relation of rational agents, as abstracted by game theory, to human flesh and blood than I can attempt here, but I can at least offer a concluding view of backward induction.

V. CONCLUSION

Rational agents, as initially defined, are indeed trapped by backward induction, in my view. Common knowledge is indeed an example of extrapolating to a limiting case that turns out not to make sense. But the trouble does not lie in the logic of the concept of knowledge, since even the omniscient can perform a *reductio ad absurdum* and the trouble is still there, if common belief is specified instead. The obstacle to rational cooperation lies in utility numbers like those got in the penny game by letting each player be motivated solely by his own gains. This is a natural approach, given a Humean view of what makes us tick, and a tempting one, granted that we have other sentiments besides greed. But once the numbers are in place and known to both players, and once they are treated as irresistible magnets, there is no escape.

Alternatively, however, Jack and Jill can have backward-looking reasons for action, alien to *A* and *B*. These reasons are not sentiments but normative features of a situation that they have partly inherited and partly created. Jack and Jill are *persons* and thus open to reflection and experience, because they are related in ways also alien to *A* and *B*. That clears the line for a long and complex comparison of Kantian with Wittgensteinian moral psychology, in the hope of deciding which has the truer account of rational agency.[6]

Game theory, meanwhile, retains its fascination but becomes more provisional. It applies in full wherever pay-offs in goods can be treated almost as if they were in utils; and much can be understood about social interaction through this style of analysis. But it cannot dispel the paradoxes that arise by its own resources. Economics is finally no more a self-contained discipline than the economy is an isolable realm of social life.

MARTIN HOLLIS

University of East Anglia

---

[6] I have tried to set the stage for this comparison in "Moves and Motives in the Games We Play," *Analysis*, L (1990): 49–62; and in "Honour Among Thieves," *Proceedings of the British Academy* (1990): 163–80. Further discussion of moral psychology can be found in chs. 5–6 of *The Cunning of Reason* (New York: Cambridge, 1988).

BOOK REVIEWS

Aristotle's First Principles. T. H. IRWIN. New York: Oxford University Press, 1988. xviii, 702 p. Cloth $75.00.

T. H. Irwin's book presents itself as an attempt to show that Aristotle was, in a certain sense, a systematic philosopher. This view of Aristotle has been widely rejected in recent scholarship for two reasons: first, the naïve assumption of ancient and medieval commentators, that Aristotle's philosophy was based upon a fixed set of fundamental principles in such a way that every part of it must be consistent with every other and had to be understood in terms of those basic principles, seems untenable in the face of numerous conflicts between the doctrines of different treatises. Especially since developmental studies became popular from the 19th century on, it seemed much more plausible to explain such inconsistencies as signs of change or development over the nearly forty years of Aristotle's career as a philosophical writer. Secondly, Aristotle's own remarks (in the *Posterior Analytics*) about the separateness of different branches of science explicitly discourage the assumption that his treatises might be unified by a common systematic background. Irwin does not wish to return to the old naïve view, of course, but he argues that with the appearance of "first philosophy" in the *Metaphysics* Aristotle develops, in contrast with his earlier methodology, a set of metaphysical "first principles" that underlie many of his later specialized treatises. The title of Irwin's book, then, should be understood in two ways. First we get an examination of Aristotle's (early) views on first principles (Part I), then an exposition of the doctrines from the *Metaphysics* which serve as first principles in the later works (Part II). The third part supports the other two by case studies meant to show how Aristotle's metaphysical doctrines are at the basis of his theories of the soul, of ethics, and of politics.

Much of the interest of the book lies in this third part, which offers detailed analyses of the main arguments of the *Ethics* and the *Politics*. One cannot quite help thinking that the epistemological and methodological theme was to some extent superimposed on these studies. And that is just as well, for if one finds the arguments of the first two parts less than convincing, the third will still be worth studying for its incisive and rigorous analyses of crucial doctrines.

Since we have virtually no information about the relative chronology of Aristotle's works independent of their contents, an interpretation that argues for development and changes of mind must be

0022-362X/91/8809/489–96

cautious not to beg any questions by assuming the chronology that best fits the interpretation. Irwin is well aware of this point, and his chronological assumptions are properly modest. He supposes, with most modern scholars, that the treatises traditionally grouped together as the *Organon* are earlier than most of the other extant works. Other treatises are then considered as earlier or later depending on their closeness or otherwise to the philosophical positions of the *Organon*. Here disagreement about interpretation will obviously invalidate chronological conclusions but, as Irwin rightly points out toward the end (472), most of his argument is in fact independent of chronological considerations. Development, as I think most scholars would now agree, offers the most plausible explanation of the difference between the *Categories* and *Posterior Analytics*, on the one hand, and the *Metaphysics*, on the other, and this seems to be the only crucial assumption.

The main argument of parts I and II is roughly as follows. In the early treatises, Aristotle puts forth a set of epistemological and metaphysical views which leads to serious difficulties about the status of his own philosophical doctrines. On the one hand, Aristotle assumes metaphysical realism and epistemological foundationalism; on the other hand, his account of the philosophical investigation of first principles, i.e., basic assumptions of scientific knowledge, seems to show that all he can attain by his method of "dialectic" is a theory that achieves coherence among "common beliefs" (*endoxa*),[1] which will not support a claim to objective truth. In the *Posterior Analytics*, Aristotle offers an intuitionist account of the knowledge of first principles; but apart from being very sketchily stated, this theory is open to fatal objections. Aristotle seems to have no defense against the charge that he cannot distinguish between the correct grasp of a first principle and the mere belief that one has grasped a first principle. His method might suggest that principles must be grasped by inference to the best explanation, but this is incompatible with Aristotle's own assertion that knowledge of first principles cannot be based on inference. Hence Aristotle's own theories of substance, essence, change, and so on seem to rest on no more than what Irwin calls "pure dialectic," an argument that is based on more or less widely held beliefs. Philosophers who do not give much weight to common opinions could therefore simply reject Aristotle's theories as irrelevant. But it seems clear that Aristotle himself presents his views as objective truths, not just common opinion.

---

[1] I would prefer to follow Jonathan Barnes in translating this term as "reputable beliefs," since the views considered include those of other philosophers which may be far removed from common opinions. According to Aristotle, it is the orator, not the dialectician, who has to limit himself to widely accepted beliefs, given the audience he is trying to persuade. But see Irwin, 494, n. 42.

Since intuitionism does not seem to offer a way out of the dilemma, Aristotle revises his methodology and, beginning in the *Metaphysics,* introduces what Irwin calls "strong dialectic," a method that builds upon a subset of common beliefs and relies only on those which an opponent could not seriously refuse to accept. The paradigm case of an argument to show that certain propositions are such that one cannot seriously deny them is Aristotle's defense of the principle of noncontradiction in *Met.* iv. 4. Aristotle argues that anyone who claims to reject the principle will find himself assuming its truth in the very attempt to state its denial. Irwin connects this and the subsequent antiskeptical arguments in *Met.* iv. 4–6 with the *Metaphysics*' project of a most general science of being, claiming that they together establish that anyone who wishes to make a serious claim to knowledge will have to assume the existence of persistent subjects with essences. These, then, will be the proper objects of first philosophy, and in *Met.* vii–ix Aristotle offers a revised theory of substance, built upon more solid foundations than the "pure dialectic" of the *Categories.* The new metaphysical doctrine comes to form the foundation of later treatises, such as the *De Anima,* which argues that the soul is the form and substance of the body, and the "political science" expounded in the *Ethics* and the *Politics,* which support their doctrine of the human good by reference to Aristotle's metaphysical views on substances and natural teleology.

While I would agree with Irwin about the contrasts between the *Categories* and the *Posterior Analytics* as opposed to the *Metaphysics,* I do not think that the hypothesis of a methodological change of mind, or a transition from "pure" to "strong" dialectic is warranted or needed to explain Aristotle's development. Obviously, a revision of an earlier doctrine can be prompted by many considerations, and the changes between the *Metaphysics* and the *Organon* seem to me to be sufficiently motivated by the need to make room for the new concepts of matter and form, on the one hand, and by a consideration of the status of philosophical theorizing with regard to the special sciences, on the other. In the *Topics*—Aristotle's official handbook of dialectic—Aristotle says that dialectic has a "way towards the principles of all disciplines" (*Top.* 1, 101b3) in virtue of its capacity to examine propositions of all kinds. He does not say, however, as Irwin seems to assume, that it is also dialectic that discovers the first principles by offering a theory that will accommodate or explain the "reputable beliefs" from which an investigation must begin. Theory building is the task of the philosopher, and Aristotle's theories seem to conform to this model—he accepts a theory if it succeeds in preserving or justifying the most widely held or authoritative beliefs and explaining the rejected ones by showing how they arise from understandable errors. Now, according to Irwin, such coherence will not suffice to satisfy the demands of Aristotle's objectivist epistemology, for there could, of course, be several different theories that meet the requirement of coherence. Furthermore, since not all of the reputable beliefs can be true, it is possible that a proposed theory is coherent, but accepts the wrong subset of common beliefs. But it is not clear to me that Aristotle would have been aware of the first point and, as to the second, one might appeal to Aristotle's epistemological optimism, expressed repeatedly in the statement that the opinions of all reasonable people must contain some grain of truth (see, e.g., EE i.6, 1216b28-31; EN i.8, 1098b27-9). This is why he insists that a good theory should not only preserve as many as possible of the reputable beliefs, but also try to explain those which are rejected, on the ground that, if they are not true, they must at least be understandable misapprehensions of the truth, not mere madness and fantasy. Given this optimistic view, Aristotle's position in attacking Plato or Democritus for their contempt of common beliefs is stronger than Irwin admits. Finally, instead of diagnosing a rejection of the intuitionist account of the grasp of first principles in the *Posterior Analytics,* we might fit this into the methodological picture by supposing that we come to grasp the basic principles of a science by realizing that they explain the beliefs from which we had started. Irwin thinks that Aristotle must reject such a move because he maintains that first principles are grasped "noninferentially," and the fact that a certain set of principles offers the best explanation would clearly constitute a reason for accepting them, rather than mere intuitive insight. But I do not think that Aristotle rules out this kind of reason. When he argues that the first principles of a science must be "indemonstrable" and hence grasped without proof, he merely excludes the possibility of deductive demonstration, and an argument to show that we have good reason to believe that p need not be an argument to show that p is the case, let alone an explanation of why it is the case that p.

So far I see no reason to think that Aristotle was caught in a methodological dilemma. The methodology just described seems to be the same as what we find in later treatises. What is new in the *Metaphysics* is the thesis that there can be a discipline, called "wisdom" or "first philosophy," that is most general and hence prior to all the specialized sciences, and deals with being qua being. The *Posterior Analytics* had denied the possibility of such a super science on the ground that the first principles of each science are independent and not derivable from any other science. In the *Metaphysics,* Aristotle declares that such a science is possible after all, and that it also deals with the so-called "common principles," i.e., propositions like the principle of noncontradiction that are accepted by each of the special sciences. We do not know, since Aristotle does not tell us, why Aristotle changed his mind about the possibility of a most general discipline, but it is a fair guess that he might have come to see that his own ontological doctrines, put forward with great confidence, for example, in the *Categories,* did not fit into the picture outlined in the *Posterior Analytics.* However that may be, it seems clear that the investigation of first principles cannot be the task of a demonstrative science, since every such science must assume its axioms and cannot justify them. So Aristotle has to resort to the kind of "negative demonstration" offered in defense of the principle of noncontradiction in *Met.* iv, or indeed to his old method, dialectical examination of reputable views and construction of a theory that will either justify them or show them to be understandable errors. It seems worth noting that the "negative demonstration" of *Met.* iv. 4 is rather untypical for Aristotle and not (as far as I can see) paralleled elsewhere in the *Metaphysics.* On the other hand, it seems to be inspired by Plato's arguments against Protagoras in the *Theaetetus.* Being, like Plato's arguments, mainly a refutation (first of people who pretend to deny the principle of noncontradiction, then of skeptics about knowledge), the series of arguments in *Met.* iv. 4–6 is certainly not officially presented as establishing a fundamental principle about possible objects of knowledge. (The Kantian association is intended here: see 540, n. 18.)

Hence it seems rather bold to assume that this principle—namely, that there must be persistent subjects with essences—established by "strong dialectic," serves as the basis of Aristotle's subsequent disquisitions about substance in *Met.* vii–ix. As far as the modest results of bk.iv will go, a Platonist, or someone who dislikes multiplying entities beyond necessity, could perfectly well shrug off Aristotle's immensely complicated efforts to find a candidate for the role of substance (basic entity) which will be both a basic subject and an essence. Irwin's succinct account of Aristotle's line of argument is an admirable tour de force (no doubt it will be hotly debated by the numerous experts in this particular field) which seems persuasive to the extent that it shows how Aristotle might have arrived at his view (if it is indeed his considered view) that particular forms are the best candidates for the contested position. But given the large number of strange premises (the most mystifying, to my mind, is that "basic subjects must be identical to their essences") that need to be added to reach this result, it can hardly count as simply derived from the epistemological arguments of *Met.* iv. The discussion of substance in the *Metaphysics* is certainly far more sophisticated than the one in the *Categories,* as it has to be once the distinction between matter and form is introduced, but as far as I can see, its epistemological foundations are the same.

Irwin's second illustration of the gap between pure and strong dialectic seems to me entirely unconvincing. Irwin claims that the first two books of the *Physics* are a good example of the weaknesses of pure dialectic in that Aristotle works only with common beliefs, and hence fails to see the deeper implications of his own doctrine of the four "causes." I think that the doctrine is not as muddled as Irwin makes it out to be, and so it is not surprising that it is found in exactly the same version in the—allegedly more advanced—*Metaphysics.*

According to Irwin, Aristotle involves himself in an inconsistency, succinctly described as follows:

> Aristotle seems to accept this inconsistent set: (1) Form and matter are origins of change. (2) The origin of change is the same as the efficient cause. (3) Therefore form and matter are efficient causes. (4) The formal and material causes are not efficient causes. (5) Therefore form and matter are not efficient causes. (6) Therefore they are not origins of change. To defend (1) he must reject (2) or (4) or both (517f, fn. 1).

But (2) is not Aristotle's view, and hence there is no inconsistency. Irwin thinks that (2) must be Aristotelian because Aristotle describes the matter and form of a natural thing as "origin (or principle—the Greek word is the same) of change," and also the efficient cause (*Phys.* ii. 3, 195a11 = *Met.* v. 2, 1013b11; the standard formula is slightly different). It does not follow, however, that all origins or principles of change are efficient causes. The simplest way to show this is presumably by appeal to *Met.* v. 1, the chapter that lists different uses of the term *arche* ("origin" or "principle"). There Aristotle explicitly distinguishes between internal and external principles (of change, as his examples show). To clinch the point, the external principle is illustrated both with the standard formula that introduces the efficient cause and with the standard examples: parent/child, insult/battle. Form and matter are, of course, internal principles of change (of the thing constituted by them), and hence they are not efficient causes (of the same thing). That does not preclude the form from being the efficient cause in relation to another thing: as Aristotle often remarks, the form *in the parent* is the efficient cause of the offspring; which shows, incidentally, that Irwin's (5) does not follow from (4). In general, it is as mistaken to assume that "form" means the same as "formal cause" as it is to think that "principle of change" means the same as "efficient cause." When Aristotle says (*Phys.* ii. 7, 198a25-27) that in natural things formal, final, and efficient cause often coincide, he means that the form has three different explanatory roles, not that there are really only two kinds of causes. There is therefore no need to attempt a reduction of the four causes to the efficient cause, nor to speculate, as Irwin does, that Aristotle has overlooked the crucial role of the efficient cause because he is paying too much attention to common beliefs. If Irwin were right, Aristotle's doctrine would conspicuously fail to cover his own paradigm case of scientific demonstration through explanatory premises, namely mathematical proof. There is no room for material, final, or efficient causes in mathematics, since mathematics deals with immaterial and unchanging entities; the only sort of "cause" in this field must be formal, as exemplified in such propositions as that a triangle has angles equal to 180° *because* it is a triangle.

There seems then to be no good reason to explain Aristotle's development by appeal to alleged revisions in his methodology. Presumably dialectic was never the only method of philosophy, since it does not include theory building, and the epistemological foundations of the search for first principles may well have remained the same throughout, except for the introduction of "first philosophy."

Now, fortunately, one does not need to agree with the developmental hypothesis to accept the main thesis of Irwin's part III, namely, that Aristotle uses his metaphysical doctrines of substance, nature, teleology, and so on in many of his other works. That he does so seems overwhelmingly evident in the *De Anima,* but it is also true, I think, of the *Nicomachean Ethics* and the *Politics,* where Irwin supports his case by showing how Aristotle's theory of the human good relies on his natural teleology. One might only wish to modify Irwin's claim by saying that the metaphysical principles could come from the *Physics* just as well as from the *Metaphysics*.

Irwin's extended analysis of Aristotle's "political science" (ethics and politics taken together) is penetrating and often illuminating. In the *Ethics,* he pursues the question whether Aristotle can justify his apparent assumption that human virtue, conceived as perfection of rational agency, is identical with the set of commonly recognized moral virtues. In the *Politics,* he shows in several cases that, if one pays attention to the difference between basic principles and empirical assumptions, one can see that some of the rather appalling conclusions do not follow from the principles, and hence provide no reason to reject the entire theory as ideologically biased. I would not agree with Irwin's answer to the first question, but it would take too long to argue this in the space of an already long review, and the reader should look for herself.

What remains of the thesis that Aristotle was a systematic philosopher? If it is sufficient for this claim that Aristotle should be using the same metaphysical framework in many of his works, wherever it is relevant, one might find the thesis rather uncontroversial. But Irwin clearly wishes to establish more—namely, that Aristotle consciously developed his metaphysical doctrines as a foundation for his subsequent theories, and that he himself considered them as the basic framework within which one should evaluate and understand the more specialized theories. (The relation would be similar to that between Kant's *Critique of Pure Reason* and his moral philosophy.) If this were true, then one should expect Aristotle to pay far more attention than he actually does to the implications of his metaphysical premises within the respective special fields. As Irwin admits, Aristotle never even explicitly states the—legitimate and important—question Irwin raises about the relation between virtue as perfection of rational agency and the traditional moral virtues. Hence one suspects that Aristotle himself, though freely taking over propositions from his metaphysics where he could use them, did not see those "first principles" as a kind of theoretical backbone that ought to structure the rest of his discussion. Perhaps he ought to have been a systematic philosopher; but I very much doubt that he tried to be one. Whether this was a fault or a blessing will have to be decided case by case.

As far as Irwin's book is concerned, I think that the parts turn out to be far more interesting than the whole. Irwin's interpretations are always worth reading and thinking about, whether one agrees with him or not, and he is surely right in saying that the sort of discussion he offers is needed if one wants to take Aristotle seriously as a philosopher and not just treat him as a museum piece. No scholar with a serious interest in Aristotle can afford to ignore this book.

Harvard University GISELA STRIKER

What Reason Demands. RÜDIGER BITTNER. Theodore Talbot, trans. New York: Cambridge University Press, 1989. ix + 198 p. Cloth $39.50, paper $13.95.

This book is a revised and expanded edition of Rüdiger Bittner's *Moralisches Gebot oder Autonomie*.[1] It is divided into seven chapters, each beginning with a concise summary of its contents. The first chapter sharply poses and explains the main issue: Does reason demand that one be moral? The next three chapters criticize contemporary attempts to establish an affirmative answer. The fifth chapter maintains that Kant's principle of autonomy, suitably modified, provides a necessary and sufficient condition for moral laws, but it concludes moral laws are "null and void," "practically irrelevant" (106, 107). The remaining chapters develop and defend the radical thesis that what reason demands is only prudence, not morality.

Bittner begins by arguing that the crucial question for moral philosophy is not "Why should I be moral?" but rather "Should I be moral?" Moral demands are construed as familiar kinds of utterances, like insults and greetings, and so what is in doubt is not whether moral demands exist but whether they are "justified" or "valid." Bittner distinguishes between justifying a demand and demonstrating its validity. To justify a demand is to give reasons *why* one ought to heed it, that is, to say what is good about doing as prescribed. To prove the validity of a demand we need only to give appropriate reasons *that* one should conform, and the sort of reasons that are relevant will depend on the kind of demand in question. A demand that one present one's automobile for inspection, for example, may be justified by the fact that conformity will improve road safety, but the validity of the demand can be shown by citing regulations for vehicles enacted by an authority with jurisdiction.

"Why should I be moral?" asks whether moral demands are justified, and "Should I be moral?" asks whether they are valid. Philosophers have failed to give convincing answers to the first question, according to Bittner, and so it is best to focus on the second. Here Bittner concedes that moral demands could be "valid" even though "unjustifiable." That is, conceivably we could have reasons for thinking *that* we should heed moral demands even though we lack reasons that tell us *why* (in the sense that they explain what is "desirable" about heeding such demands).

Although to show the validity of moral demands one would not

---

[1] Freiburg: Alper, 1983.

0022-362X/91/8809/497–501

need to provide "justifying" reasons, one would still need to state reasons of some kind appropriate for moral demands. Unlike the demand for automobile inspection, moral demands obviously cannot be shown to be valid by citing official regulations. As commonly understood, moral demands implicitly make strong claims that severely constrain what could count as a demonstration of their validity. To show that a certain moral demand is valid for a person, then, it is not enough to point out that the person has adequate reasons *of some kind* to do what is demanded. Merely prudential reasons, for example, would not validate the demand *as moral;* and, on Bittner's view, even if all rational agents would agree on moral rules prescribing the action, that would not validate the prescription *as a demand* on people actually unwilling to comply.

In chapters II–IV, Bittner tries to show that the validity of moral demands cannot be based on essential human attributes or accomplishments, actual or hypothetical contracts, or promises. The argument here proceeds primarily by way of a critical review of the views of contemporary writers, including Kurt Baier, Neil Cooper, Philippa Foot, Alan Gewirth, Geoffrey Russell Grice, R.M. Hare, A.I. Melden, Thomas Nagel, Jan Narveson, Richard Norman, John Rawls, David A.J. Richards, and John Searle. Bittner finds none that provides adequate grounds for accepting the validity of moral demands. His objections are crisply stated and pertinent, though often familiar and too brief to be convincing without more discussion.

The wide-ranging scope of these critical chapters should be evident from the following summary of the main points. Gewirth mistakenly assumes that regarding something good entails claiming a right to it. Norman wrongly infers the validity of moral demands from the claim that actions can be understood only by moral concepts. Nagel confuses the weak conclusion that one can meaningfully extend practical judgments from one's own case to others with the stronger claim that the extended judgments are valid. What individuals in Rawls's original position would agree upon is hard to determine, and in any case agreement among fictitious individuals would give us no reason to accept the corresponding moral demands as valid. Searle's argument from the supposed analyticity of "Promises ought to be kept" has been undermined by counterarguments from Hare and others. Prichard relies on an unintelligible notion that one can put oneself under obligation, or bind oneself irrevocably, by one's own action; but in "promising" one can really only communicate intentions and so, strictly speaking, there are no promises.

Finding no satisfactory arguments in contemporary sources, Bittner turns back to Kant. No philosopher, of course, could disagree with Bittner's final conclusions more than Kant, but, as Bittner recognizes, Kant himself forcefully articulated the challenge to moral philosophy that Bittner believes cannot be met. The crucial elements are contained in the second chapter of the *Groundwork*. First, analysis reveals that in judging that an act is a duty one implies that the act is inescapably required by reason, independently of prudential considerations. Second, such rational requirements are conceived and experienced as internal constraints on one's will to which one cannot be indifferent; and so these requirements cannot be equated with conventional rules, the commands of others, or any natural facts. Third, one would be subject to such constraints if but only if they were self-imposed, that is, willed for oneself by oneself as a rational agent with autonomy. These conceptual points, Kant thought, pose a challenge that previous moral philosophy never adequately faced: What grounds are there for believing, and indeed how could it be, that there are constraints on our choices that are (in the relevant senses) rational, not optional, and yet self-imposed? If the challenge cannot be met, Kant insists, the conclusion must be that ordinary moral judgment rests on a false presupposition and so duty is a mere illusion.

Kant, of course, believed that he gave an adequate response to this challenge in the third chapter of the *Groundwork,* or at least in the first part of the *Critique of Practical Reason*. Kant's arguments may be unconvincing to contemporary moral philosophers, but most manage to avoid the skeptical conclusion that there are no duties by rejecting one or another of Kant's initial presuppositions. Bittner, interestingly, does not reject the basic ideas behind Kant's challenge but instead argues that, suitably interpreted and modified, the problem Kant posed is real, that Kant failed to resolve it, and that indeed it cannot be resolved. Bittner's conclusion, surprisingly, is not that there are no valid moral *laws* but rather that there are no valid moral *demands*. A modified Kantian principle of autonomy states necessary and sufficient conditions for moral laws to be valid for a person, and these conditions may sometimes be met. But moral laws cannot support moral demands (alleged "duties" and "obligations") because they have no rational claim on those who do not want to act on them. For those willing to comply, moral rules impose no *demands;* for those unwilling to comply, they are not *valid*.

Imagining a child tempted not to share candy with his brother and unmoved by preliminary answers, Bittner gives the ultimate response that encapsulates his view: "To the child's question 'Why should I be moral?' our answer was this: No reasons can be seen. To the new question, 'Should I be moral?', we now have the answer: If you want to be, yes; otherwise, no" (107).

For those willing to accept enough of the Kantian background to acknowledge prima facie force in Kant's challenge, the crucial steps in Bittner's argument will be his interpretation of autonomy and his modified Kantian autonomy principle. Autonomy, on Bittner's view, is not a duty or an ideal; it is simply a property of agency manifested whenever we act intentionally, according to our ideas and plans. His principle of autonomy states that moral rules are valid moral laws for a person if and only if both (1) all rational beings could agree to the rules and (2) *the person in question actually adopts the rules as guides for his own conduct.*

The last clause, which is the key to Bittner's unKantian conclusions, is supposed to be a refinement of Kant's idea that moral laws are "self-imposed." Kant clearly did not intend the autonomy condition as Bittner understands it, for Kant did not suppose that one could render a moral rule invalid in his case simply by wanting and choosing to disregard it. Bittner realizes, of course, that his autonomy principle is a modification of Kant. What he offers in support of this crucial principle is primarily a critique of Kant. The argument, in effect, proceeds by first assuming an initial plausibility in the Kantian idea that autonomy is necessary for moral laws, next interpreting and criticizing Kant's attempts to articulate the appropriate autonomy condition, and then accepting the residue (however implausible) as true. The argument may be challenged at each stage, but especially dubious, I think, is the move from "self-imposed laws" to whatever fits with an agent's intentions. If Kant's metaphor of self-legislation has any initial plausibility at all, it does so only under an interpretation quite incompatible with the idea that one "legislates" to oneself whatever one wants. Any interpretation, or modification, of Kant's idea of autonomy that reduces self-legislation this far undercuts the initial grounds for supposing that autonomy is a necessary condition for moral validity.

In a brief review one can do little more than explain an author's project, main conclusions, and strategy of argument, but *What Reason Demands,* in my opinion, merits more sustained critical attention. Bittner's discussions of Kant and others are informed and challenging even when one disagrees, and he presents the case for his own radical thesis forcefully and concisely. The last two chapters of his book are constructive and rich with fresh ideas on prudence, developing the theme that "practical reason is not a judge, but a physician" (128). In all, despite substantial disagreements, I think that moral philosophers ought to read Bittner's book. That is, reason advises this, even if it cannot demand it.

THOMAS E. HILL, JR.
University of North Carolina/Chapel Hill

NOTES AND NEWS

The editors regret to announce the untimely death of Alan H. Donagan, the Doris and Henry Dreyfuss Professor of Philosophy at the California Institute of Technology, on Wednesday, May 29. Professor Donagan joined the faculty at Caltech in 1984, after spending 14 years at the University of Chicago, where he was the Phyllis Fay Horton Professor of Humanities. Previously he was a professor at the University of Illinois/Urbana-Champaign, Indiana University, and the University of Minnesota. At both Indiana and Minnesota, he served as chairman of the philosophy department. Earlier, he held academic appointments in philosophy at University College in Canberra, Australia, and at the University of Western Australia in Perth.

Behavior and Philosophy, a journal for the interdisciplinary philosophical study of psychology and behavioral science published by the Cambridge Center for Behavioral Studies, is pleased to announce that Max Hocutt will succeed George Graham as editor, commencing with Volume XX, 1992. All correspondence concerning the journal, including manuscripts for submission, may be sent to Max Hocutt, Phil. Dept., Univ. of Alabama, University, AL 35487.

Stanford Humanities Center will offer up to six external fellowships for 1992–93 intended for postdoctoral scholars and teachers in the humanities, or those in other fields working on related projects, who are interested in spending the academic year at Stanford. The fellowships are primarily intended to enable fellows to pursue their own research and writing; however, recipients are also expected to devote about one-sixth of their time to teaching or in some other way contributing to intellectual life at Stanford. The deadline for application is November 15. Application materials and further information may be obtained by writing to Fellowship Program, Stanford Humanities Center, Mariposa House, Stanford University, Stanford, CA 94305-8630.

NEW BOOKS

ADAMS, E. M.: *The Metaphysics of Self and World: Toward a Humanistic Philosophy*. Philadelphia: Temple, 1991. xv, 325 p. Cloth $54.95.

AVNI, ORA: *The Resistance of Reference: Linguistics, Philosophy, and the Literary Text*. Baltimore: Johns Hopkins, 1990. xv, 313 p. Cloth $39.95.

BROADIE, ALEXANDER: *The Tradition of Scottish Philosophy: A New Perspective on the Enlightenment*. Savage, MD: Barnes & Noble, 1990. vii, 144 p. Cloth $42.00.

CARME, STEWART ZANE: *Vulgarity and Authenticity: Dimensions of Otherness in the World of Jean-Paul Sartre*. Amherst: Massachusetts UP, 1991. x, 255 p. Cloth $29.95.

CODE, LORRAINE: *What Can She Know? Feminist Theory and the Construction of Knowledge*. Ithaca: Cornell, 1991. xiv, 349 p. Cloth $42.50, paper $14.95.

CRANSTON, MAURICE: *The Noble Savage: Jean-Jacques Rousseau 1754–1762*. Chicago: University Press, 1991. xiv, 399 p. Cloth $32.50.

CRITTENDEN, CHARLES: *Unreality: The Metaphysics of Fictional Objects*. Ithaca: Cornell, 1991. xiii, 177 p. Cloth $26.50, paper $9.95.

CURRIE, GREGORY: *The Nature of Fiction*. New York: Cambridge, 1990. xii, 222 p. Cloth $39.50.

DORE, CLEMENT: *Moral Scepticism*. New York: St. Martin's, 1991. x, 137 p. Cloth $45.00.

DUFF, R. A.: *Intention, Agency, and Criminal Liability: Philosophy of Action and the Criminal Law*. Cambridge: Blackwell, 1990. xiv, 219 p. Cloth $44.95, paper $15.95.

DUMMETT, MICHAEL: *Frege and Other Philosophers*. New York: Oxford, 1991. xiv, 330 p. Cloth $69.00.

EVANS, J. CLAUDE: *Strategies of Deconstruction: Derrida and the Myth of the Voice*. Minneapolis: Minnesota UP, 1991. xxiii, 205 p. Cloth $39.95, paper $14.95.

FONTANA, BIANCAMARIA: *Benjamin Constant and the Post-Revolutionary Mind*. New Haven: Yale, 1991. xvii, 165 p. Cloth $25.00.

FRIEDMAN, WILLIAM J.: *About Time: Inventing the Fourth Dimension*. Cambridge: MIT, 1990. x, 147 p. Cloth $19.95.

GALSTON, MIRIAM: *Politics and Excellence: The Political Philosophy of Alfarabi*. Princeton: University Press, 1990. xi, 240 p. Cloth $35.00.

GOODMAN, RUSSELL B.: *American Philosophy and the Romantic Tradition*. New York: Cambridge, 1991. xi, 162 p. Cloth $34.50.

GOULD, THOMAS: *The Ancient Quarrel between Poetry and Philosophy*. Princeton: University Press, 1990. xxvii, 318 p. Cloth $39.50.

GRIFFIN, NICHOLAS: *Russell's Idealist Apprenticeship*. New York: Oxford, 1991. xiii, 410 p. Cloth $98.00.

HANNAY, ALASTAIR: *Human Consciousness*. New York: Routledge, 1990. x, 221 p. Cloth $47.50.

HARDING, SANDRA: *Whose Science? Whose Knowledge? Thinking from Women's Lives*. Ithaca: Cornell, 1991. xii, 319 p. Cloth $34.50, paper $12.95.

HARRIS, ZELLIG: *A Theory of Language and Information: A Mathematical Approach*. New York: Oxford, 1991. xii, 428 p. Cloth $69.00.

HILL, CHRISTOPHER S.: *Sensations: A Defense of Type Materialism*. New York: Cambridge, 1991. ix, 253 p. Cloth $44.50, paper $15.95.

HUNLEY, J. D.: *The Life and Thought of Friedrich Engels: A Reinterpretation*. New Haven: Yale, 1991. xiii, 184 p. Cloth $22.50.

ISHIGURO, HIDE: *Leibnitz's Philosophy of Logic and Language*. 2nd ed. New York: Cambridge, 1991. x, 246 p. Cloth $49.50, paper $16.95.

JOHNSON, LAWRENCE E.: *A Morally Deep World: An Essay on Moral Significance and Environmental Ethics*. New York: Cambridge, 1991. ix, 301 p. Cloth $39.50.

KADISH, MORTIMER R.: *Toward an Ethic of Higher Education*. Stanford: University Press, 1991. ix, 205 p. Cloth $29.50.

KYMLICKA, WILL: *Liberalism, Community, and Culture*. New York: Oxford, 1991. 280 p. Paper $19.95.

LEFTOW, BRIAN: *Time and Eternity*. Ithaca: Cornell, 1991. xii, 377 p. Cloth $42.95.

LEONARD, STEPHEN T.: *Critical Theory in Political Practice*. Princeton: University Press, 1990. xxiv, 295 p. Cloth $35.00.

LUCAS, J. R., and P. E. HODGSON: *Spacetime and Electromagnetism: An Essay on the Philosophy of the Special Theory of Relativity*. New York: Oxford, 1990. xv, 310 p. Cloth $63.00, paper $29.95.

MCCARTHY, THOMAS: *Ideals and Illusions: On Reconstruction and Deconstruction in Contemporary Critical Theory*. Cambridge: MIT, 1991. x, 254 p. Cloth $22.50.

MILLAR, ALAN: *Reasons and Experience*. New York: Oxford, 1991. x, 227 p. Cloth $69.00.

NIELSEN, KAI: *After the Demise of the Tradition: Rorty, Critical Theory, and the Fate of Philosophy*. Boulder, CO: Westview, 1991. x, 278 p. Cloth $38.50.

NORTON, DAVID L.: *Democracy and Moral Development: A Politics of Virtue*. Berkeley: California UP, 1991. xvi, 198 p. Cloth $29.95.

PERNER, JOSEF: *Understanding the Representational Mind*. Cambridge: MIT, 1991. xiv, 348 p. Cloth $35.00.

PHILLIPS, ANNE: *Engendering Democracy*. University Park: Pennsylvania State UP, 1991. 183 p. Cloth $28.50, paper $13.95.

PLATTS, MARK: *Moral Realities: An Essay in Philosophical Psychology*. New York: Routledge, 1991. viii, 232 p. Cloth $74.50.

PYBUS, ELIZABETH: *Human Goodness: Generosity and Courage*. Toronto: University Press, 1991. x, 149 p. Cloth $40.00.

QUINE, W. V., and RUDOLF CARNAP: *Dear Carnap, Dear Van: The Quine-Carnap Correspondence and Related Work*. RICHARD CREATH, ed. Berkeley: California UP, 1991. xv, 484 p. Cloth $34.95.

REGAN, TOM: *The Thee Generation: Reflections on the Coming Revolution*. Philadelphia: Temple, 1991. xi, 180 p. Cloth $24.95.

ROWE, WILLIAM L.: *Thomas Reid on Freedom and Morality*. Ithaca: Cornell, 1991. x, 189 p. Cloth $26.95.

SCHLESINGER, GEORGE: *The Sweep of Probability*. Notre Dame: University Press, 1991. viii, 228 p. Cloth $28.95.

SMITH, P. CHRISTOPHER: *Hermeneutics and Human Finitude: Toward a Theory of Ethical Understanding*. New York: Fordham, 1991. xxviii, 291 p. Cloth $40.00.

SNARE, FRANCIS: *Morals, Motivation, and Convention: Hume's Influential Doctrines*. New York: Cambridge, 1991. xii, 322 p. Cloth $44.50.

SORELL, TOM: *Scientism: Philosophy and the Infatuation with Science*. New York: Routledge, 1991. xi, 206 p. Cloth $45.00.

SOSA, ERNEST: *Knowledge in Perspective: Selected Essays in Epistemology*. New York: Cambridge, 1991. xi, 250 p. Cloth $49.50, paper $16.95.

SPRAGENS, THOMAS A.: *Reason and Democracy*. Durham: Duke, 1990. xi, 281 p. Cloth $39.50, paper $17.95.

TILGHMAN, B. R.: *Wittgenstein, Ethics and Aesthetics: The View from Eternity*. Albany: SUNY, 1991. xiv, 193 p. Cloth $49.50, paper $16.95.

WEATHERFORD, ROY: *The Implications of Determinism*. New York: Routledge, 1991. x, 275 p. Cloth $45.00.

WESTBROOK, ROBERT B.: *John Dewey and American Democracy*. Ithaca: Cornell, 1991. xix, 570 p. Cloth $29.95.

WETTSTEIN, HOWARD: *Has Semantics Rested on a Mistake? And Other Essays*. Stanford: University Press, 1991. xi, 232 p. Cloth $29.50.

YOLTON, JOHN W.: *Locke and French Materialism*. New York: Oxford, 1991. vii, 239 p. Cloth $55.00.

ZUIDERVAART, LAMBERT: *Adorno's Aesthetic Theory: The Redemption of Illusion*. Cambridge: MIT, 1991. xxiv, 380 p. Cloth $35.00.

Philosophy of Science

Gaston Bachelard, Subversive Humanist
Texts and Readings
MARY McALLESTER JONES

McAllester Jones provides a generous sampling of Bachelard's seminal writings about science, poetry, and human consciousness. She frames her translations with critical essays on the development of Bachelard's ideas and his significant relationship to the contemporary French intellectual revolution commonly associated with Michel Foucault and Jacques Derrida.

Science and Literature series—George Levine, General Editor

Paper $15.00 Library cloth edition $37.50

The University of Wisconsin Press
114 North Murray Street • Madison, WI 53715 • (608) 262-2994
Wisconsin

110146

110146

Compiled
1999-2000